MANAGEMENT SCIENCE
Articles in Advance, pp. 1–17
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

http://dx.doi.org/10.1287/mnsc.2015.2158

The Structural Virality of Online Diffusion

Sharad Goel, Ashton Anderson
Stanford University, Stanford, California, 94305 {scgoel@stanford.edu, ashton@cs.stanford.edu}

Jake Hofman, Duncan J. Watts
Microsoft Research, New York, New York 10016 {jmh@microsoft.com, duncan@microsoft.com}

Viral products and ideas are intuitively understood to grow through a person-to-person diffusion process analogous to the spread of an infectious disease; however, until recently it has been prohibitively difficult to directly observe purportedly viral events, and thus to rigorously quantify or characterize their structural properties. Here we propose a formal measure of what we label "structural virality" that interpolates between two conceptual extremes: content that gains its popularity through a single, large broadcast and that which grows through multiple generations with any one individual directly responsible for only a fraction of the total adoption. We use this notion of structural virality to analyze a unique data set of a billion diffusion events on Twitter, including the propagation of news stories, videos, images, and petitions. We find that across all domains and all sizes of events, online diffusion is characterized by surprising structural diversity; that is, popular events regularly grow via both broadcast and viral mechanisms, as well as essentially all conceivable combinations of the two. Nevertheless, we find that structural virality is typically low, and remains so independent of size, suggesting that popularity is largely driven by the size of the largest broadcast. Finally, we attempt to replicate these findings with a model of contagion characterized by a low infection rate spreading on a scale-free network. We find that although several of our empirical findings are consistent with such a model, it fails to replicate the observed diversity of structural virality, thereby suggesting new directions for future modeling efforts.

Keywords: Twitter; diffusion; viral media
History: Received August 14, 2013; accepted November 26, 2014, by Lorin Hitt, information systems.
Published online in *Articles in Advance*.

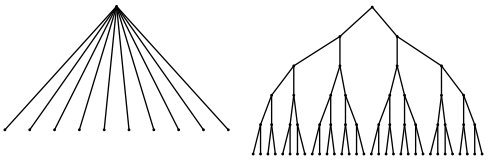
1. Introduction

When a piece of online media content—say, a video, an image, or a news article—is said to have "gone viral," it is generally understood not only to have rapidly become popular, but also to have attained its popularity through some process of person-to-person contagion, analogous to the spread of a biological virus (Anderson and May 1991). In many theoretical models of adoption (Coleman et al. 1957, Bass 1969, Mahajan and Peterson 1985, Valente 1995, Bass 2004, Toole et al. 2012), in fact, this analogy is made explicit: an "infectious agent"—whether an idea, a product, or a behavior—is assumed to spread from "infectives" (those who have it) to "susceptibles" (those who do not) via some contact process, where susceptibles can then be infected with some probability.[1] Both intuitively and also in formal theoretical models, therefore, the notion of viral spreading implies a rapid, large-scale increase in adoption that is driven largely, if not exclusively, by peer-to-peer spreading. Clearly, however, viral spreading is not the only mechanism by which a piece of content can spread to reach a large population. In particular, mass media or marketing efforts rely on what might be termed a "broadcast" mechanism, meaning simply that a large number of individuals can receive the information directly from the same source. As with viral events, broadcasts can be extremely large—the Superbowl attracts over 100 million viewers, while the front pages of the most popular news websites attract a similar number of daily visitors—and hence the mere observation that something is popular, or even that it became so rapidly, is not sufficient to establish that it spread in a manner that resembles social contagion.

Figure 1 schematically illustrates these two stylized modes of distribution—broadcast and viral—where the former is dominated by a large burst of adoptions from a single parent node, and the latter comprises a multigenerational branching process in which any one node directly "infects" only a few others. Although the stylized patterns in Figure 1 are intuitively plausible and also easily distinguishable from one another, differentiating systematically

[1] Even models of social contagion that do not correspond precisely to the mechanics of biological infectious disease (for example, "threshold models" (Granovetter 1978) make different assumptions regarding the nonindependence of sequential contacts with infectives (Lopez-Pintado and Watts 2008)) assume some form of person-to-person spread (Watts 2002, Kempe et al. 2003, Dodds and Watts 2004).

Figure 1 **A Schematic Depiction of Broadcast vs. Viral Diffusion, Where Nodes Represent Individual Adoptions and Edges Indicate Who Adopted from Whom**



between broadcast and viral diffusion requires one, in effect, to characterize the fine-grained structure of viral diffusion events. Yet, in spite of a large theoretical and empirical literature on the diffusion of information and products, relatively little is known about their structural properties, in part because the requisite data have not been available until very recently, and in part because the concept of virality itself has not been formulated previously in an explicitly structural manner. Classical diffusion studies (Coleman et al. 1957, Rogers 1962, Bass 1969, Valente 1995, Young 2009, Iyengar et al. 2010), for example, typically had access to only aggregate diffusion data, such as the cumulative number of adoptions of a product, technology, or idea over time (Fichman 1992). In such cases, the observation of an S-shaped adoption curve—indicating a period of rapid growth followed by saturation—is typically interpreted as evidence of social contagion (Rogers 1962); however, S-shaped adoption curves may also arise from broadcast distribution mechanisms such as marketing or mass media (Van den Bulte and Lilien 2001). Compounding the difficulty, real diffusion events are unlikely to conform precisely to either of these conceptual extremes. In a highly heterogeneous media environment (Walther et al. 2010, Wu et al. 2011), where any given piece of content can spread via email, blogs, and social networking sites as well as via more traditional offline media channels, one would expect that popular content might have benefited from some possibly complicated combination of broadcasts and interpersonal spreading.

To understand the underlying structure of an event, therefore, one must reconstruct the full adoption cascade, which in turn requires observing both individual-level adoption decisions and also the social ties over which these adoptions spread. Only recently have data satisfying these requirements become available, as a result of online behavior such as blogging (Adar and Adamic 2005, Yang and Leskovec 2010), e-commerce (Leskovec et al. 2006), multiplayer gaming (Bakshy et al. 2009), and social networking (Sun et al. 2009, Yang and Counts 2010, Bakshy et al. 2011, Petrovic et al. 2011, Goel et al. 2012, Hoang and Lim 2012, Tsur and Rappoport 2012, Kupavskii et al. 2012, Jenders et al. 2013, Ma et al. 2013).

A second empirical challenge in measuring the structure of diffusion events, which has in fact been highlighted by these recent studies, is that the vast majority of cascades—over 99%—are tiny and terminate within a single generation (Goel et al. 2012). Large and potentially viral cascades are therefore necessarily very rare events; hence, one must observe a correspondingly large number of events to find just one popular example, and many times that number to observe many such events. As we will describe later, in fact, even moderately popular events occur in our data at a rate of only about one in a thousand, whereas "viral hits" appear at a rate closer to one in a million. Consequently, to obtain a representative sample of a few hundred viral hits—arguably just large enough to estimate statistical patterns reliably—one requires an initial sample on the order of a billion events, an extraordinary data requirement that is difficult to satisfy even with contemporary data sources.

In this paper, we make three distinct but related contributions to the understanding of the structure of online diffusion events. First, we introduce a rigorous definition of *structural virality* that quantifies the intuitive distinction between broadcast and viral diffusion and allows for interpolation between them. As we explain in more detail below, our definition is couched exclusively in terms of observed patterns of adoptions, not on the details of the underlying generative process. Although this approach may seem counterintuitive in light of our opening motivation (which does make reference to generative models), the benefit is that the resulting measure does not depend on any modeling assumptions or unobserved properties, and hence can be applied easily in practice. Also importantly, by treating structural virality as a continuously varying quantity, we skirt any categorical distinctions between completely "broadcast" and "viral" events, allowing instead for open-ended and fine-grained distinctions between these two extremes; that is, events can be more or less structurally viral without imposing any particular threshold for becoming or "going" viral.

Our second contribution is to apply this measure of structural virality to investigate the diffusion of nearly a billion news stories, videos, pictures, and petitions on the microblogging service Twitter. To date, most studies directly documenting person-to-person diffusion have been limited to a small set of highly viral products (Liben-Nowell and Kleinberg 2008, Dow et al. 2013), leaving open the possibility that such hand-selected events are astronomically rare and not representative of viral diffusion more generally. In contrast, by systematically exploring the structural properties of a billion events on Twitter, we aim to estimate the frequency of structurally viral cascades, quantify the diversity in the structure of cascades,

and investigate the relationship between cascade size and structure. It could be, for example, that the most popular content is also extremely viral, but equally it could be that successful products are mostly driven by mass media (i.e., a single large broadcast) or by some combination of broadcasts and word of mouth. Depending on the relative importance of broadcast versus viral diffusion in driving popularity, that is, the relationship between popularity and structural virality could be positive (larger events are dominated by viral spreading), negative (larger events are dominated by broadcasts), or neither (all events regardless of size exhibit a similar mix of broadcasts and virality, which scale together). Applying our structural virality measure to a representative sample of successful cascades, we find evidence for the third possibility, namely, that the correlation between popularity and virality is generally low. Moreover, for any given size (equivalent popularity), structural virality is extremely diverse: cascades can range between pure "broadcasts," in the sense that all adopters receive the content from the same source, and highly "viral," in the sense of comprising multigenerational branching structures.

The third contribution of this paper is to compare our empirical observations of cascade structure to predictions from a series of simple generative models of diffusion. Specifically, we conduct large-scale simulations of a simple disease-like contagion model, similar to the original Bass (1969) model of product adoption, on a network comprising 25 million nodes. In the simplest variant, we assume that the infectiousness of the "disease" is a constant, and the network on which it spreads is an Erdős–Rényi (ER) random graph. In successively more complicated variants, we allow the infectiousness to vary, or the network to be "scale free" (i.e., where the number of neighbors can vary from tens to tens of millions), or both. Because large diffusion events are so rare, we also conduct on the order of 1 billion simulations per parameter setting, necessitating over 100 billion simulations in total. We find that although our simplest models are incapable of replicating even the most general features of our empirical data, a still-simple model comprising constant infectiousness and scale-free degree distribution can capture many, but not all, of the observed features. We conclude with some suggestions for future modeling efforts.

2. Defining Structural Virality

We now turn to our first goal of defining structural virality. Before proceeding, we reemphasize that our notion of structural virality is intended to complement, not substitute for, the many existing generative models of viral propagation and their associated parameters (Bass 1969, Granovetter 1978, Watts 2002, Kempe et al. 2003, Dodds and Watts 2004). To clarify, generative models attempt to describe the underlying diffusion mechanism itself—for example, as a function of the intrinsic infectiousness of the object that is spreading, or of the properties of the contact process or the network over which the diffusion occurs, or of the timescales associated with adoption. By contrast, our notion of structural virality is concerned exclusively with characterizing the structure of the observable adoption patterns that arise from some unobserved generative process. Naturally, the particular value of structural virality associated with some event will in general depend on the underlying generative process—as indeed we will demonstrate in §5, where we introduce and study several such models. Importantly, however, our desired *definition* of structural virality should not depend on these particulars. In other words, regardless of what contagion process is (assumed to be) responsible for some piece of content spreading or what network it is spreading over, the end result is some pattern of adoptions that exhibits some structure, and our goal is to characterize a particular property of that structure.

Recalling also that our goal is to disambiguate between the broadcast and multigenerational branching schematics depicted in Figure 1, we first lay out some intuitively reasonable criteria that we would like any such metric to exhibit. First, for a fixed total number of adoptions in a cascade, structural virality should increase with the branching factor of the structure: specifically, it should be minimized for the broadcast structure on the left of Figure 1 and should be relatively large for structures with a high branching factor, as on the right of Figure 1. Second, for a fixed branching factor, structural virality should increase with the number of generations (i.e., depth) of the cascade; that is, all else equal, larger branching structures should be more structurally viral than smaller ones. Finally, and in contrast with multigenerational branching structures, larger broadcasts should not be any more structurally viral than smaller broadcasts; hence we require that, for the extreme case of a pure broadcast, structural virality be approximately independent of size.

A natural choice for such a metric is simply the number of generations, or depth, of the cascade. Indeed, after size, depth is one of the most widely reported summary statistics of diffusion cascades (Liben-Nowell and Kleinberg 2008, Goel et al. 2012, Dow et al. 2013). One problem with depth, however, is that a single, long chain can dramatically affect the measure. For example, a large broadcast with just one, long, multigenerational branch has large depth, even though we would not intuitively consider it to be structurally viral. To correct for this issue, one could

instead consider the average depth of nodes (i.e., the average distance of nodes from the root). This average depth measure alleviates the problem of a handful of nonrepresentative nodes skewing the metric, and intuitively distinguishes between broadcasts and multigenerational chains. Even this measure, however, fails in certain cases. Notably, if an idea or product traverses a long path from the root and then is broadcast out to a large group of adopters, the corresponding cascade would have high average depth (since most adopters are far from the root) even though most adoptions in this case are the result of a single influential node.

Addressing the shortcomings of both depth and average depth, we focus our attention on a classical graph property studied originally in mathematical chemistry (Wiener 1947), where it is known as the "Wiener index." Specifically, we define structural virality $\nu(T)$ as the average distance between all pairs of nodes in a diffusion tree T; that is, for $n > 1$ nodes,

$$\nu(T) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}, \tag{1}$$

where d_{ij} denotes the length of the shortest path between nodes i and j.[2] Equivalently, $\nu(T)$ is the average depth of nodes, averaged over all nodes in turn acting as a root.

Our metric $\nu(T)$ provides a continuous measure of structural virality, with higher values indicating that adopters are, on average, farther apart in the cascade, and thus suggesting an intuitively viral diffusion event. In particular, as with depth and average depth, over the set of all trees on n nodes $\nu(T)$ is minimized on the star graph (i.e., the stylized broadcast model in Figure 1) where $\nu(T) \approx 2$. Moreover, a complete k-ary tree (as in Figure 1 with $k = 2$) has structural virality approximately proportional to its height; hence, structural virality will be maximized for structures that are large and that become that way through many small branching events over many generations.[3]

Although $\nu(T)$ satisfies some basic requirements of theoretical plausibility, as with the other candidate measures we discussed it is possible to construct hypothetical examples for which the corresponding numerical values are at odds with the motivating intuition. For example, a graph comprised of two stars connected by a single, long path has large $\nu(T)$ but would not intuitively be considered viral. Whether or not such pathological cases appear with any meaningful frequency is, however, largely an empirical matter, and hence the utility of the metric must ultimately be evaluated in the context of real examples, which we discuss in detail below as well as in Appendix B.

3. Data and Methods

Our primary analysis is based on approximately 1 billion diffusion events on Twitter, where an event constitutes the independent introduction of a piece of content into the social network—including videos, images, news stories, and petitions—along with all subsequent repostings of the same item.[4] Specifically, we include in our data all tweets posted on Twitter that contained URLs pointing to one of several popular websites over a 12 month period, from July 2011 to June 2012.[5] In total, we observe roughly 622 million unique pieces of content; however, because individual pieces of content can be posted by multiple users, we observe approximately 1.2 billion "adoptions" (i.e., posting of content). Although our data are not a total sample of Web content that is shared on Twitter,[6] they do include the vast majority and hence are essentially unbiased at least with respect to Tweets linking to Web content.[7] Importantly for our conclusions, our sample also exhibits considerable diversity both with respect to production and consumption. For example, a typical online video is likely to have been produced and distributed by

[2] Naive computation of $\nu(T)$ requires $O(n^2)$ time; however, as discussed in Appendix B, a more sophisticated approach yields a linear-time algorithm (Mohar and Pisanski 1988), facilitating computation on very large cascades.

[3] Somewhat more precisely, for any branching ratio $k << n$, $\nu(T)$ increases with size n, whereas for $k \approx n$ (i.e., pure broadcasts) it does not; hence, increasing popularity corresponds to increasing structural virality only when it arises from "viral" spreading, not merely from larger broadcasts.

[4] We use the term "reposting" rather than the more conventional "retweet" because individuals frequently repost content that they receive from another user without using the explicit retweet functionality provided by Twitter, or even acknowledging the source of the content.

[5] For news those websites include bbc.co.uk, cnn.com, forbes.com, nytimes.com, online.wsj.com, guardian.co.uk, huffingtonpost.com, news.yahoo.com, usatoday.com, telegraph.co.uk, and msnbc.msn.com. For video they include youtube.com, m.youtube.com, youtu.be, vimeo.com, livestream.com, twitcam.livestream.com, ustream.tv, twitvid.com, mtv.com, and vh1.com. For images they include twitpic.com, instagr.am, instagram.com, yfrog.com, p.twimg.com, twimg.com, i.imgur.com, imgur.com, img.ly, and flickr.com. For petitions they include change.org, twitition.com, kickstarter.com.

[6] URLs and redirects were dereferenced from original tweets, and extraneous query parameters were removed from URLs to identify multiple versions of identical content. To avoid left censoring of our data (i.e., missing the initial postings of a URL), we look for occurrences of the URLs during the month prior to our analysis period and only include in our sample instances where the first observation does not appear before July 1, 2011. To avoid right censoring, we restrict to tweets introduced prior to June 30, 2012, but continue tracing the diffusion of these tweets through July 31, 2012.

[7] It is of course possible that Tweets containing links to Web content are systematically different from other Tweets in ways that might affect our conclusions. For this reason, in Appendix D we conduct a separate analysis of tweets containing long hashtags, which are unlikely to diffuse outside of Twitter, finding qualitatively similar results.

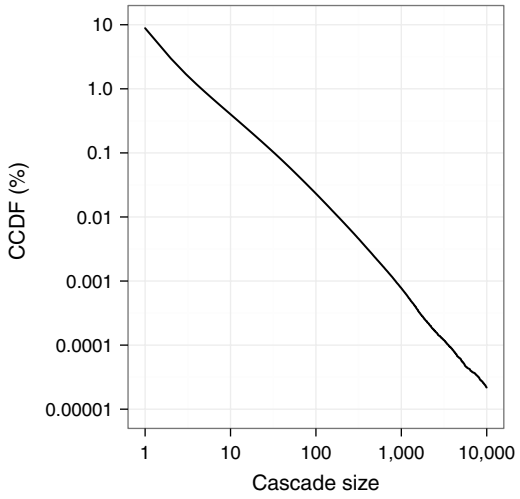
an amateur videographer uploading his or her own work onto YouTube, whereas an article appearing in a major news outlet was likely written by a professional reporter. Moreover, the experience of watching a video is quite distinct from that of reading a news article, both in terms of the time and effort required on the part of the consumer and also their goals—for example, to be entertained versus informed—in doing so. Due in part to these qualitative differences on both the supply and also demand sides of the market for media, we find large quantitative differences in the frequency of the four domains; specifically, images and videos are far more numerous than news stories, and petitions are by far the least numerous. For similar reasons, therefore, one might also expect qualitatively distinct sharing mechanisms to dominate in different domains, leading to different patterns both of popularity and also structural virality.

To evaluate the structure of online diffusion, for each independent introduction of a unique piece of content in our data we construct a corresponding diffusion "tree" that traces each adoption back to a single "root" node, namely, the user who introduced that particular piece of content.[8] Specifically, for each observation of a URL whose diffusion we seek to trace, we record (1) the adopter (i.e., the identity of the user who posted the content); (2) the adoption time (i.e., the time at which the content was posted); and (3) the identities of all users the adopter follows—hereafter referred to as the adopter's "friends"—from whom the adopter could conceivably have learned about the content. For each such event, we first determine whether at least one of the adopter's friends adopted the same piece of content previously. If no such friend exists, then the adopter is labeled a "root" of the resulting diffusion tree; otherwise, the friend who adopted the content most recently before the focal adopter—and who is most likely to have exposed the focal user to the content—is labeled the focal adopter's "parent." Although there is at times genuine ambiguity in determining the proximate cause of an adoption, in many cases adopters explicitly credit another individual in their tweet, allowing us to accurately infer an adopter's parent in approximately 95% of instances (see Appendix C for details of the tree construction algorithm and the associated evaluation procedure).

4. Results

Consistent with previous work (Bakshy et al. 2011, Goel et al. 2012), we find that the average size of these diffusion trees (also referred to interchangeably as "cascades" or "diffusion events") is 1.3—meaning that for every 10 introductions of content, there are on average three additional downstream adoptions. More strikingly, and as noted in Goel et al. (2012), we also find that the vast majority of cascades terminate within a single generation; specifically, about 99% of adoptions are accounted for either by the root nodes themselves or by the immediate followers of root nodes. As noted previously (Goel et al. 2012), however, the preponderance of small and shallow events does not rule out the possibility that large, structurally interesting events do occur, only that they occur sufficiently infrequently so as not to be observed even in relatively large data sets. Exploiting the fact that we have a much larger data set than in previous studies—over a billion observations in our initial sample—we therefore now focus exclusively on the subsample of rare events that qualify as large, and hence have the potential to be structurally interesting. Specifically, hereafter we restrict attention to the 0.025% of diffusion trees containing at least 100 nodes (Figure 2), a requirement that leaves us with roughly 1 out of every 4,000 cascades, and thus reduces the number of cascades we study in detail from approximately 1 billion to 219,855.

Figure 2 **Distribution of Cascade Sizes on a Log–Log Scale, Aggregated Across the Four Domains We Study: Videos, News, Pictures, and Petitions**



Note. CCDF = complementary cumulative distribution function.

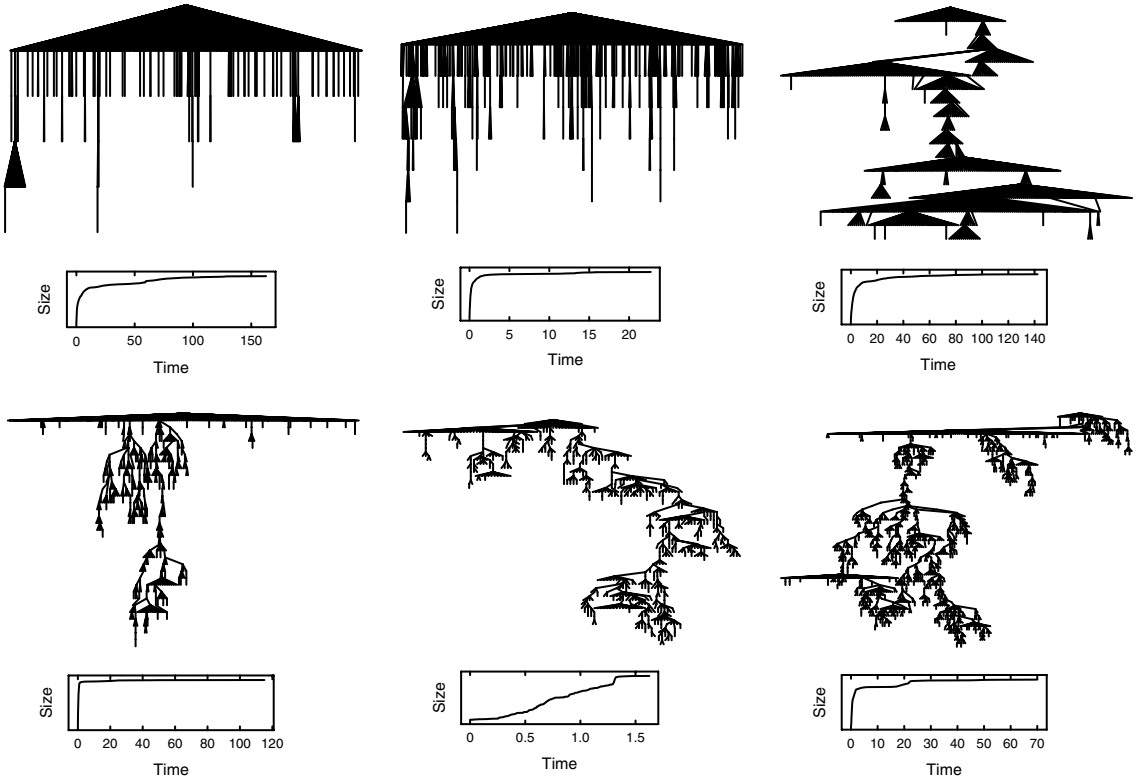
4.1. Structural Diversity

From this subpopulation of "successful" diffusion events, Figure 3 displays a stratified random sample ordered by structural virality $\nu(T)$. Specifically, cascades with between 100 and 1,000 adopters were ranked by $\nu(T)$ and logarithmically binned, and a random cascade was then drawn from each bin.[9] We

[8] Although diffusion trees are in reality dynamic objects, meaning that they grow over time as new adoptions take place, here we treat them as static objects representing the final state of a given diffusion process.

[9] We note that this exercise was performed only once to avoid hand selection of the best "random" sample.

Figure 3 A Random Sample of Cascades Stratified and Ordered by Increasing Structural Virality, Ranging from 2 to 50

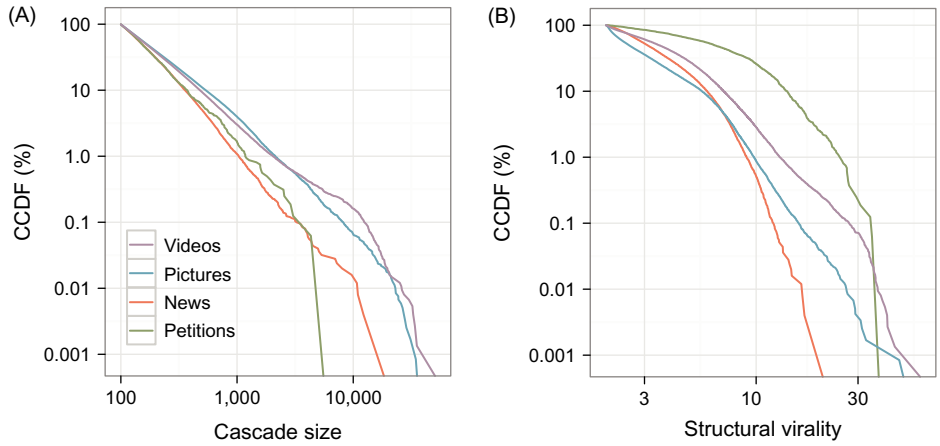


Notes. For ease of visualization, cascades were restricted to having between 100 and 1,000 adopters. Cumulative adoption curves (i.e., total cascade size over time) are shown below each cascade, with time indicated in hours. For visual clarity, the adoption curves terminate at 99% of the final cascade size.

observe that the ordering from left to right and top to bottom by increasing $\nu(T)$ is strikingly consistent with how these same structures would be ranked intuitively in order of increasing virality, not only in the trivial case of disambiguating broadcast and viral extremes, but also in making relatively fine-grained distinctions between intermediate cases. Thus, $\nu(T)$ not only seems to be a reasonable measure of structural virality in theory, but also performs well in practice. Considering now the cumulative adoption curves shown below each cascade in Figure 3, we make two further observations. First, although the shape of these adoption curves varies considerably, from events that experience a phase of rapid growth before leveling off to events that grow almost linearly over time, there is no consistent relationship with structural virality. Strikingly, in fact, the least structurally viral of all our sampled events (top left) exhibits a cumulative adoption curve that is almost indistinguishable in shape from the most structurally viral (bottom right). Second, the timescales on which the adoptions take place (noted in hours on the horizontal axis of the cumulative plots) also varies widely, from less than an hour (bottom left) to three days (top left). As with the shape of the curves, however, there is no consistent relationship between the timescale (speed) of an adoption process and its associated structural virality. We conclude that our measure of structural virality not only effectively quantifies differences in the underlying cascade structures, but is clearly doing so by using features of the diffusion process that are not captured by aggregated data.

The ordering also highlights our first main empirical finding: Although the structures in Figure 3 are all of similar size (i.e., have similar aggregate numbers of adopters), they exhibit remarkable diversity in structure, from an approximately pure broadcast ($\nu(T) \approx 2$, top left) to an ideal-type branching structure ($\nu(T) = 34$, bottom right), with numerous intermediate variations in between. The classical literature on diffusion often posits a critical threshold—or "tipping point"—for virality, suggesting a sharp break between cascades that are viral and those that are not. If the tipping point intuition is correct, one would expect that relatively large diffusion events such as those captured in the $n = 100$ (roughly one event in 4,000) to $n = 1{,}000$ (one in 100,000) range would be dominated either by broadcasts on the one hand or by viral spreading on the other hand, but that combinations of the two should not arise. More generally, one might expect only a handful of canonical forms to account for the majority of large events: for example, some events spread exclusively via broadcast, whereas others spread exclusively via word of

Figure 4 **Size and Structural Virality Distributions on a Log–Log Scale for Cascades Containing at Least 100 Adopters, Separated by Domain**



Note. CCDF, complementary cumulative distribution function.

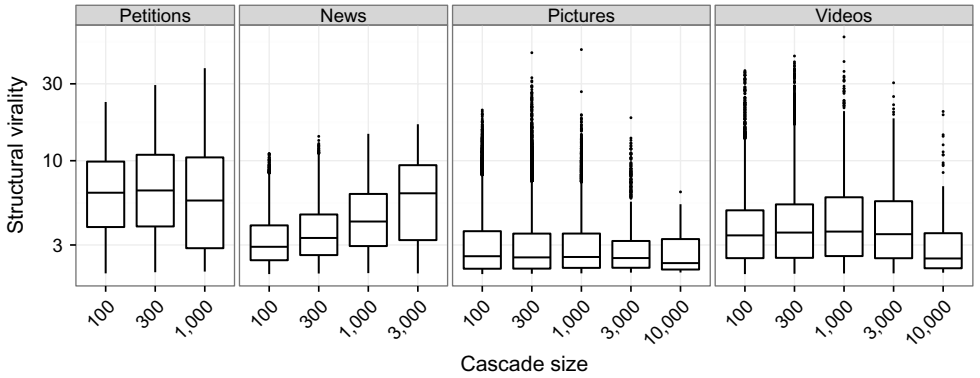
mouth, and others still spread by some combination of the two. In other words, whatever one's intuitive mental model of diffusion, one would likely expect to find that successful diffusion events of a given size would be typified by some combination of broadcast and viral diffusion, or at least some small taxonomy of types. It is striking, therefore, that Figure 3 shows examples of fine-grained variations in structural virality across the entire range of possibilities.

4.2. Examining Popularity and Structural Virality

Although Figure 3 shows that one can find examples of cascades across the spectrum of structural virality, it says little about their relative frequency or how that varies by domain. To address these questions, Figure 4(A) shows the size distribution of cascades larger than 100 adopters for all four domains—news, videos, images, and petitions—while Figure 4(B) shows the corresponding distributions of structural virality. As anticipated, Figure 4(A) shows that cascades can grow very large: For images and videos, the largest cascades attract several tens of thousands of reposts, whereas the most popular news stories are somewhat smaller (roughly 10,000 reposts), and petitions smaller still (several thousand reposts). In other words, although the vast majority of cascades are indeed small, large cascades do occur, albeit with low frequencies. Moreover, the size distributions appear to cluster into two categories: one comprising images and videos and the other comprising the rather less popular categories of petitions and news stories. In other words, the most popular videos and images are more popular than the most popular news stories and petitions not only because there are many more of the former, but also because the corresponding distributions exhibit a shallower slope; that is, for any given percentile of the relevant population, videos and images are more popular than petitions and news stories. Although we lack a compelling explanation for this systematic difference, we note that the vast majority of the most popular Twitter accounts belong not to news organizations or petition sites, but to celebrities, whose postings often contain images and videos. Moreover, YouTube and Instagram are among the top 10 most followed accounts, further facilitating the visibility of videos and images, respectively. It thus seems likely that one of the primary drivers of large image and video cascades is their promotion by individuals with large numbers of followers, consistent with past results (Bakshy et al. 2011).

Next, Figure 4(B) confirms the impression from Figure 3 that structural virality varies widely, from 2 (pure broadcast) to over 30. In particular, in contrast to classical "tipping point" theories of diffusion, we do not see a bimodal distribution of structural virality corresponding to broadcasts on the one hand and viral spreading on the other, but rather a continuous distribution of structural virality, confirming our earlier speculation that in some sense every conceivable combination of broadcasts and word-of-mouth transmission is represented. Interestingly, however, popular petitions are substantially more structurally viral than any other type of content, followed by videos, images, and news stories. For example, whereas about a quarter of popular petitions have structural virality of at least 10—meaning that petitions having garnered at least 100 adopters are quite likely to have grown virally—only about 3% of videos, 1% of images, and 0.5% of news stories exhibit the same level of structural virality. In spite of the diversity evident both in Figure 3 and Figure 4(B), therefore, the relatively larger size of cascades involving videos and images combined with their relatively low structural virality suggests that the largest cascades in those categories

Figure 5 **Box Plot of Structural Virality by Size on a Log–Log Scale, Separated by Domain**



Note. Lines inside the boxes indicate median structural virality, whereas the boxes themselves show interquartile ranges.

are not especially viral in a structural sense. In the next section, we examine this possibility in more detail.

4.3. Relationship Between Popularity and Structural Virality

As pointed out earlier, the relationship between popularity (cascade size) and structural virality is not a priori obvious; that is, depending on the empirically observed preponderance of broadcasts in small versus large events, the relationship could be positive (large events are less likely to be dominated by broadcasts than small events), negative (large events are more likely to be dominated by broadcasts than small events), or neither. Put another way, if cascades typically grow via person-to-person diffusion, we would expect structural virality to increase with cascade size. On the other hand, if large cascades are the product of broadcasts attributable to popular users on Twitter—the most popular of whom have tens of millions of followers—structural virality may not vary significantly with size, or could even decrease.

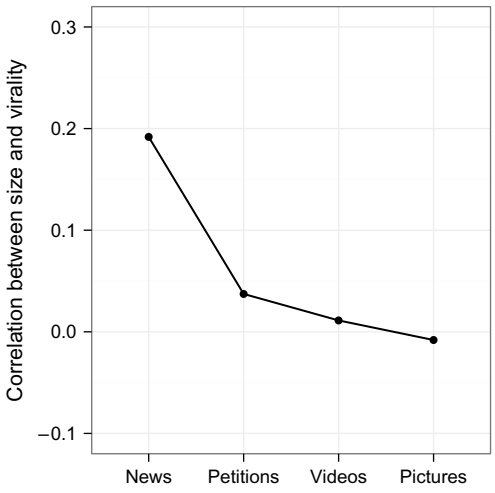
We investigate this question by examining the distribution of structural virality conditional on cascade size for each domain. First, and consistent with Figure 4, Figure 5 shows that across all sizes for which they occur, popular petitions are considerably more viral than the other domains. Second, Figure 5 shows that across all domains and size ranges, structural diversity varies considerably, confirming again the visual impression of Figure 3. Third, however, Figure 5 shows that for three out of four domains—petitions, images, and videos—median structural virality remains surprisingly invariant with respect to size. For images and videos, moreover, it is also surprisingly low: even the very largest cascades, comprising 10,000 reposts or more, exhibit median structural virality of less than 3, barely more than the theoretical minimum of 2. For petitions, meanwhile, median structural virality is between 7 and 8, roughly equivalent to a branching tree of depth between three and four generations: not a pure broadcast but still relatively shallow. Finally, for news, the relationship between size and structural virality is more positive than for the other domains, but also still surprisingly low. For cascades of size 100, for example, median structural virality is approximately 3, whereas for the largest observed news cascades, comprising 3,000 reposts, median structural virality is still less than 8, comparable to petitions.

We emphasize that there is nothing inevitable about this result. It could have been, for example, that the very largest events are characterized by multigenerational branching structures—indeed that is the clear implication of the phrase "going viral." So it is surprising that even the very largest events are, on average, dominated by broadcasts. It is also surprising that the correlation between size and structural virality is so low. As shown in Figure 6, the correlation for news is 0.2, indicating a positive but noisy relationship, whereas for petitions it is even lower (0.04), indicating almost no relationship at all, and for pictures and videos it is essentially zero. In contrast with our earlier result on diversity, which suggests that simply knowing the size of a cascade reveals very little about its structure, the combination of generally low values of structural virality and low correlation with size suggests that if popularity is consistently related to any one feature, it is the size of the largest broadcast.[10]

As in our discussion of Figure 4, we can only speculate about why (a) petitions are so much more structurally viral for every size category than other domains and (b) news stories show higher correlation between size and structural virality. We suspect, however, that the main driving factor is once again a relative dearth of large broadcast channels for petitions in particular and to a lesser extent news organizations.

[10] We also note that these results are not affected by the fact that the range of $\nu(T)$ varies with cascade size; the results are qualitatively identical when we use a measure of structural virality with a constant bounded range (see Appendix B).

Figure 6 **Correlation Between Cascade Size (Popularity) and Structural Virality Across Four Domains**



The popularity of images and videos, by contrast, is likely driven by celebrities, who increasingly have tens of millions of followers on Twitter, and whose posting behavior likely favors content of a personal and often visual nature over news and calls to action.

5. Theoretical Modeling

To recap, we have three main empirical findings. First, and consistent with previous work (Goel et al. 2012), the vast majority of diffusion events are small and accordingly lack much structure. Second, rare events that do become large exhibit striking structural diversity. And third, the size of these cascades is at most weakly correlated to their structural virality. Together these findings present an interesting theoretical question, namely, can they be replicated by a single underlying generative mechanism? And if so, what features are required? Although replicating some empirical results with a theoretical model does not on its own imply that the model is an accurate representation of the true generative process (Ijiri et al. 1977), it is nevertheless possible to rule some models out.

To address this question, we consider a series of variations on the SIR model, a classical model of biological contagion (Kermack and McKendrick 1927, Anderson and May 1991) that has frequently been adapted to model social diffusion processes,[11] initially to the specific context of new product adoption, where it is known as the Bass (1969) model, and subsequently to a wide range of other contexts including the propagation of links over a network of blogs (Leskovec et al. 2007). In any such model, there are two key sets of parameters. First, when an individual is infected (in the present case, with a piece of content), he or she subsequently infects each of his or her susceptible (i.e., not yet infected) contacts independently with probability β. Often β is assumed to be a constant, but in the current context—where it refers to the "infectiousness" of content—it is natural to think of it as being drawn from some distribution (which itself may be described by additional parameters). And second, we must specify the nature of the contact process, which here we model as a network in which $\bar{k}$ is the average node degree (i.e., the number of opportunities a typical node has to infect others) and σ^2 is the degree variance.[12]

Before proceeding, it is helpful to introduce the quantity $r = \bar{k}\beta$ (known in mathematical epidemiology as the "basic reproduction number" or R_0 of a disease). As alluded to earlier, a standard result for diseases spreading on random networks is that the condition $r = c$, where $c = 1/(1 + (\sigma/\bar{k})^2) \leq 1$, constitutes a critical threshold or tipping point, separating two regimes: a "supercritical," or "viral," regime $r > c$, in which small seeds can trigger exponential growth leading to large epidemics, and a "subcritical" regime $r < c$, in which the contagion almost surely dies out after infecting only a small number of susceptibles. From this general result, moreover, two more specific results follow. First, in Erdős–Rényi random networks $G(n, p)$, where the expected degree is $k \sim np$ and $\sigma^2 \sim k$ (as $n \to \infty$), the epidemic threshold condition reduces to $r \sim 1$ for $k \gg 1$. And second, in scale-free random networks (Barabási and Albert 1999) for which the variance diverges with the size of the network, it reduces to $r \sim 0$ as $n \to \infty$ (Pastor-Satorras and Vespignani 2001, Lyons 2000, Lloyd and May 2001), meaning that in sufficiently large scale-free networks, the subcritical regime effectively disappears.

These results are relevant to our analysis for two reasons. First, because viral events for which $r > 1$ exhibit exponential growth regardless of network structure and because we know from our data that large events are extremely rare, we restrict our analysis to the region $0 < r < 1$, corresponding to what in everyday usage would be thought of as "subcritical" spreading. Second, because we will consider both ER and scale-free random networks, the usual super-

[11] Reflecting its origins in mathematical epidemiology, the model is named for the three states—"susceptible," "infectious," and "recovered"—that each node in the network can occupy. Numerous variations of the basic SIR model have also been proposed, included the SI model, the SEIR model (where the "E" indicates "exposed"), the SIRS model, and so on (Anderson and May 1991). Here we refer to all such models canonically as SIR models.

[12] Additional parameters are also natural. For example, we only consider strict SIR models in the sense that after one time step, infected nodes are "removed" from the dynamics, meaning that they can no longer infect others nor become reinfected. Although natural for our case, where having "adopted" piece of content one cannot unadopt it, other assumptions are clearly possible, in which case additional parameters would be needed.

versus subcritical distinction is somewhat misleading. Specifically, whereas it does have a clear meaning for ER networks, for which only contagions with $r > 1$ are viral in the everyday sense of growing exponentially, in scale-free networks, all contagions are viral in the technical sense of exceeding the epidemic threshold, even though they are "dying out" as they attempt to spread.[13] As we will show next, in fact, models invoking ER networks are easily dismissed as incompatible with our empirical results, suggesting that the popular tipping point notion is largely irrelevant to the kind of viral events we study here.

We consider four models of increasing complexity and verisimilitude. In all cases, each realization of the simulation commences with an entirely susceptible population comprising 25 million individuals within which a single individual is randomly chosen to be the initially infected "seed" and proceeds until no further infections can take place.[14] We start by investigating contagions characterized by constant β spreading on an ER random graph. In light of the enormous attention paid to variations of this model both in the mathematical epidemiology (Kermack and McKendrick 1927, Anderson and May 1991) and marketing (Bass 1969, Valente 1995, Bass 2004) literatures, it is the natural baseline to consider. As noted above, however, its relevance to our empirical data can quickly be dismissed by showing that, consistent with standard theoretical results (Anderson and May 1991), the cascade size distribution is tightly centered around its mean regardless of the average network degree or infection rate, which is qualitatively different than the heavy-tailed size distribution we observe in the data.

One explanation for this result is that our assumption of constant β is unlikely to be correct. Presumably, content introduced to Twitter exhibits large differences in intrinsic interestingness and breadth of appeal, and therefore likelihood of being shared. This observation motivates the next model we consider, where the infection is again modeled as spreading on an ER graph, but the infectiousness of each piece of content, β_i, is now drawn from a power law distribution $\Pr(\beta_i) \sim \beta_i^{-\alpha}$, expressing the more plausible assumption that a large number of items in our sample are of low "quality" or "appeal" and hence are unlikely to spread (low β), whereas a small minority of appealing or high-quality items are much more likely to spread (high β). Studying this case, we do indeed recover the heavy-tailed size distribution from our empirical results. Interestingly, however, across parameter settings we consistently observe high correlation between cascade size and structural virality—because large cascades in ER must necessarily be multigenerational—which again stands in stark contrast to our empirical results. We therefore conclude that it is the ER network, not necessarily the assumption about constant item quality, that is responsible for the poor model fit.

Thus motivated, we now examine a third model in which we again assume β to be a constant, but the network is now a scale-free random network (Barabási and Albert 1999), constructed using the configuration method[15] (Newman 2005, Clauset et al. 2009), reflecting the roughly power law degree distribution $p(k) \sim k^{-\alpha}$ observed for Twitter (Bakshy et al. 2011). Sweeping over the two parameters, α and β, we simulated content of varying infectiousness diffusing over networks with varying degree skew. Figure 7 shows the results of nearly 100 billion simulations, with 1 billion cascades generated for each parameter setting (α, β), roughly congruent with the number of cascades we analyzed on Twitter. Figure 7 shows that for certain parameters—$r \approx 0.5$ and $\alpha \approx 2.3$—the model recapitulates several important features of our empirical data.[16] First, Figure 7(A) shows that for this parameter setting the probability of a given piece of content becoming "popular"—meaning that it attracts at least 100 adoptions—is consistent with the observed rate of roughly one in one thousand. Second, Figure 7(B) shows that the mean structural virality for these parameters is 5, which again is in line with our observations. Third, Figure 7(C) shows that the correlation between size and structural virality is also in the observed range. Finally, Figure 8 shows the full marginal distributions of size and virality, and the distribution of virality conditional on

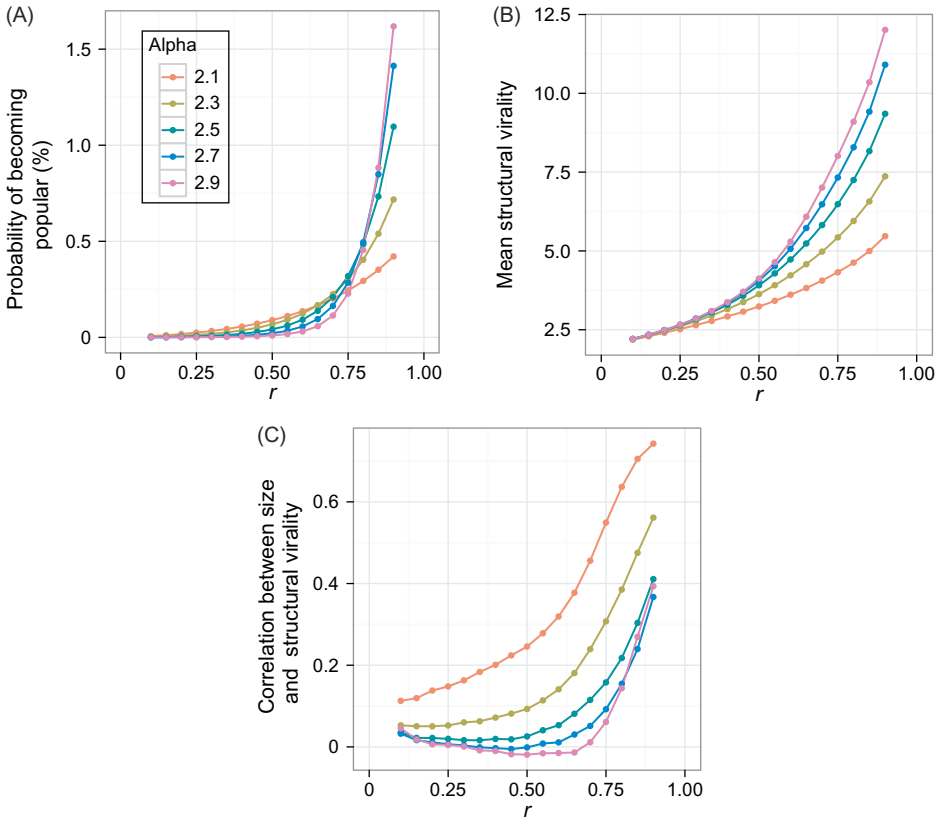
[13] The intuitive explanation for this counterintuitive result is that in scale-free networks, a typical node is likely to be connected via at most a short path to a "hub" node with an extremely high degree that, if infected, can sustain an infection that would ordinarily die out (Pastor-Satorras and Vespignani 2001).

[14] Clearly on Twitter a single unique piece of content can be introduced many times independently. In such cases, there is potential for two cascades to "collide," which clearly cannot happen in our simulations, where we introduce only one seed at a time. In light of the extreme rarity of large cascades, however, and the large size of the Twitter network, such collisions are also rare; hence, we do not believe this simplification has any significant consequences. We also note that our model is a special case of what has been called "simple contagion" (Centola 2010), in which the infection probably is independent across multiple exposures. In contrast with "complex contagion," such as occurs in "threshold models" (Granovetter 1978), where multiple exposures can combine in highly nonlinear ways, the use of individual seeds for simple contagion is relatively unproblematic.

[15] For each node in the network, its number of followers (i.e., out-degree) was first randomly selected according to a discrete power law degree distribution with exponent α, a minimum value of 10, and a maximum value of 1 million. Then nodes in the networks were randomly connected while preserving the specified degrees.

[16] The power law exponent of $\alpha \approx 2.3$ is consistent with the observed degree distribution on Twitter (Kwak et al. 2010).

Figure 7 **Likelihood of Becoming Popular (i.e., Having at Least 100 Adopters), Mean Structural Virality, and the Correlation Between Size and Structural Virality for Simulated Cascades Generated from an SIR Model on a Random Scale-Free Network, Plotted as a Function of the Model Parameters**



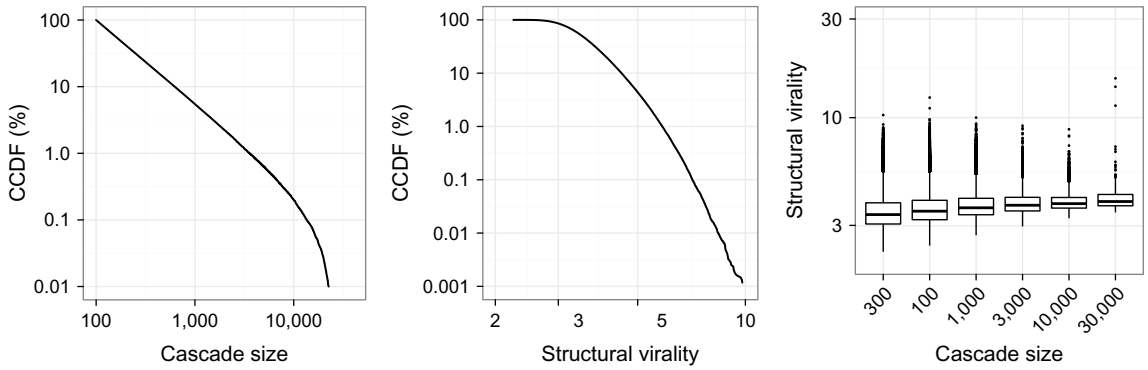
Note. Each line corresponds to a different exponent α for the power-law network degree distribution, and $r = \beta\bar{k}$ is the expected number of individuals a random node infects in a fully susceptible population.

size for this parameter choice, where we again see that the simulated cascades are similar to the empirically observed events. One notable difference between empirical and simulation results, however, is that the variance in each bin (as measured by the interquartile range) in the rightmost plot in Figure 8 is considerably less than that in Figure 5, indicating that empirical cascades exhibit much more structural diversity at any given size compared to those generated by the model.

These simulation results can be interpreted in two ways. On the one hand, it is striking that so simple a model—with only two tunable parameters—can capture many of the basic empirical regularities of what is undoubtedly a far more complex and multifaceted system. For example, although the success of real-world products is almost certainly affected by their quality, this connection is absent from our model. Indeed, for any fixed parameter choice under the SIR model, all cascades—the largest broadcasts, the most viral cascades, and the many events that acquire only a handful of adopters—have the same infectiousness β. In other words, taking infectiousness as a proxy for quality, in our simulations the largest and most viral cascades are not inherently better than those that fail to gain traction, but are simply more fortunate (Watts 2002). On the other hand, it is also interesting that our model is not able to fully capture the diversity of structural virality exhibited in the empirical data. Although we can only speculate on the reasons for this limitation, two possible explanations immediately suggest themselves. The simplest explanation is that as large as our simulated networks are (25 million nodes), they are still not as large nor is the network structure as complex as the actual Twitter follower graph, which comprises roughly 500 million users, the most connected of whom have well over 50 million followers. Possibly, therefore, the difference could be accounted for simply by increasing the size of the networks by another one or two orders of magnitude—an increase that is computationally challenging, but that is straightforward in theory. A second, and perhaps more likely, explanation is that our assumption of constant β remains too simplistic, and that introducing such variation into our model would also increase the variation of structural virality at any given size.

The fourth and final model that we simulate therefore replaces constant β with β_i drawn from a power

Figure 8 **Box Plot of Structural Virality by Size (on a Log–Log Scale) for 1 Billion Simulated Cascades Generated from an SIR Model on a Random Scale-Free Network with $\alpha = 2.3$ and $r = 0.5$**



Note. CCDF, complementary cumulative distribution function.

law distribution, identical to the ER case in our second model above. Surprisingly, however, a similarly extensive set of simulations using this model finds that it does not in fact lead to noticeably more structural diversity; moreover, it leads to high correlation between size and structural virality. The reason for both results is that higher (lower) values of β_i generate larger (smaller) events, not more (less) structurally viral events of the same size. Thus, even though the diversity of β_i does affect the size distribution of cascades, for a given cascade size it does not generate more diversity of structural virality. Identifying a mechanism that accounts for the observed diversity of structural virality therefore presents an interesting challenge for future modeling work.

6. Discussion

Returning to our opening motivation, our paper makes three main contributions. First, we have introduced the concept of structural virality, one of the first measures to formally quantify the structure of information cascades. Although our results are restricted to the diffusion of information on Twitter, our structural approach to diffusion processes applies quite generally, both to online and offline settings. It is often claimed, for example, that some of the most successful Internet products in recent history, such as Hotmail, Gmail, and Facebook, were driven primarily by word-of-mouth adoption, in part because the companies that created these products did not initially have large advertising budgets, and in part because by design they contained features to explicitly encourage sharing. Yet these products also benefitted from extensive media coverage, which might have driven large numbers of adoptions from a small number of broadcast events. Likewise, although popular Internet memes are typically described as having spread virally, they also typically receive substantial media coverage. Without reconstructing the actual sequence of events by which a given product, idea, or piece of content was adopted, and relatedly without a metric for quantifying virality, the mere observation of popularity—however rapidly accrued—allows one to conclude little about the relative importance of viral versus broadcast mechanisms in determining the observed outcome. With the appropriate data, therefore, our notion of structural virality could conceivably shed light on a much broader range of diffusion processes than we have considered here.

Our second contribution is to measure the fine-grain structure of nearly 1 billion naturally occurring diffusion events in a specific online setting, namely, Web content spreading on Twitter. In particular, we have identified hundreds of thousands of large cascades—the biggest such collection to date—revealing remarkable structural diversity of diffusion events, ranging from broadcast to viral and containing essentially everything in between, where we emphasize that such an exercise would be difficult absent a metric for classifying and ordering the structure of these cascades automatically. In addition, we find relatively low correlation between size and virality, highlighting the difficulty of determining how content spread given only knowledge of its popularity.

Third, we have shown that a simple model of contagion is broadly consistent with our empirical findings. The theoretical literature has largely focused on supercritical diffusion processes to model large, viral cascades; however, the vast majority of diffusion events comprise only a few nodes, and rarely extend beyond one generation beyond the root node, or seed (Goel et al. 2012). Events of this latter kind are naturally attributable to subcritical diffusion,[17] and hence one might thus be tempted to model online diffusion via two categorically distinct mechanisms, separately accounting for the head and tail

[17] For example, Leskovec et al. (2007) found that a susceptible-infected-susceptible (SIS) model with $\beta = 0.025$, equivalent to $r \approx 0.14$, was able to replicate the size distribution of observed cascades of links over a network of blogs.

of the distribution. Indeed, the very label "viral hit" implies precisely the exponential spreading of the sort observed in contagion models in their supercritical regime. It is therefore notable that essentially everything we observe, including the very largest and rarest events, can be accounted for by a simple model operating entirely in the low infectiousness parameter regime. Indeed our best model fit is for $r \approx 0.5$, which is considerably lower even than a previous "subcritical" estimate of $\beta \approx 0.99$ based on the diffusion of chain letters (Golub and Jackson 2010)—a difference that is likely due to the heavy-tailed (scale-free) degree distribution of Twitter.[18]

Finally, in addition to our three scientific contributions, we note that our work also contributes to the emerging field of computational social science in the sense that it addresses a traditional social science question—How does content spread via social networks?—but answers it using a type and scale of data that has only recently become available; that is, only after tracing the propagation of over a billion pieces of content can we collect an unbiased sample of large, and exceedingly rare, cascades to observe their subtle structural properties. By contrast, previous work (Goel et al. 2012) that investigated the propagation of nearly one million news stories and videos—one of the largest diffusion studies at the time—was only able to observe relatively small events, resulting in a qualitatively incomplete view of diffusion. In a similar vein, the most relevant previous analysis of the structure of extremely large diffusion events relied on just two examples, specifically the reconstructed paths of two Internet chain letters (Liben-Nowell and Kleinberg 2008). Although collecting even two such examples required considerable ingenuity, it is nevertheless the case that inferring general principles from so few observations is inherently difficult (Golub and Jackson 2010, Chierichetti et al. 2011). One of our main findings, in fact, is that large diffusion events exhibit extreme diversity of structural forms—a finding that necessarily requires many examples. Thus, although our current work is by no means exhaustive, its scale facilitates a significant step toward describing the nature and diversity of online information diffusion.

Appendix A. Computing Structural Virality

The average distance measure of structural virality that we use, $\nu(T)$, has often been applied in mathematical chemistry, where it is known as the Wiener index, and its efficient computation has also long been known. For completeness, here we present a simple and scalable method to compute $\nu(T)$. We begin by showing how the Wiener index, as well as the average depth of a tree, can be expressed in terms of the sizes of various subtrees.

Lemma 1. *For a tree T with n nodes, let* $\text{depth}_{\text{avg}}$ *denote the average depth of nodes in the tree. Letting $\mathcal{S}$ be the set of all subtrees of T, we have*

$$\frac{1}{n}\sum_{S\in\mathcal{S}}|S| = \text{depth}_{\text{avg}} + 1.$$

Proof. For any node $v_i \in T$ and any subtree $S \in \mathcal{S}$, let $\delta_S(v_i)$ be 1 if $v_i \in S$ and 0 otherwise. Then,

$$\begin{aligned}\sum_{S\in\mathcal{S}}|S| &= \sum_{S\in\mathcal{S}}\sum_{i=1}^{n}\delta_S(v_i)\\ &= \sum_{i=1}^{n}\sum_{S\in\mathcal{S}}\delta_S(v_i)\\ &= \sum_{i=1}^{n}1+\text{depth}(v_i).\end{aligned}$$

The result now follows by dividing each side by n. □

Theorem 2. *For a tree T with n nodes, let* $\text{depth}_{\text{avg}}$ *denote the average depth of nodes in the tree, let* dist_{avg} *denote the average distance between all pairs of distinct nodes (i.e.,* $\text{dist}_{\text{avg}} = \nu(T)$*), and let $\mathcal{S}$ be the set of all subtrees of T. Then,*

$$\text{dist}_{\text{avg}} = \frac{2n}{n-1}\left[1+\text{depth}_{\text{avg}} - \frac{1}{n^2}\sum_{S\in\mathcal{S}}|S|^2\right]. \tag{A1}$$

In particular,

$$\text{dist}_{\text{avg}} = \frac{2n}{n-1}\left[\frac{1}{n}\sum_{S\in\mathcal{S}}|S| - \frac{1}{n^2}\sum_{S\in\mathcal{S}}|S|^2\right]. \tag{A2}$$

Proof. Statement (A2) in the theorem follows directly from (A1) together with Lemma 1, and so we only need to establish statement (A1). For any two nodes $v_i, v_j \in T$, let $\text{LCA}(v_i, v_j)$ denote their lowest common ancestor: the unique node in T of greatest depth that has both v_i and v_j as descendants (where a node is allowed to be a descendant of itself). Since the shortest path between v_i and v_j goes through $\text{LCA}(v_i, v_j)$, we have

$$\begin{aligned}\text{dist}(v_i, v_j) &= \text{dist}(v_i, \text{LCA}(v_i, v_j)) + \text{dist}(\text{LCA}(v_i, v_j), v_j)\\ &= [\text{depth}(v_i) - \text{depth}(\text{LCA}(v_i, v_j))]\\ &\quad + [\text{depth}(v_j) - \text{depth}(\text{LCA}(v_i, v_j))]\\ &= \text{depth}(v_i) + \text{depth}(v_j) - 2\cdot\text{depth}(\text{LCA}(v_i, v_j)).\end{aligned}$$

Let $\text{subtrees}(v_i, v_j)$ be the set of subtrees that contain both v_i and v_j, and observe that this set consists of exactly those subtrees that contain $\text{LCA}(v_i, v_j)$. Since for any node v there are $1 + \text{depth}(v)$ subtrees that contain it,

$$|\text{subtrees}(v_i, v_j)| = 1 + \text{depth}(\text{LCA}(v_i, v_j)).$$

Substituting this expression into the previous equation, we see that

$$\text{dist}(v_i, v_j) = 2 + \text{depth}(v_i) + \text{depth}(v_j) - 2|\text{subtrees}(v_i, v_j)|.$$

[18] We note that this finding also recalls earlier work that sought to account for the surprisingly long-term and low-level persistence of computer viruses in terms of a low-infectiousness contagion spreading over a scale-free network (Pastor-Satorras and Vespignani 2001). Although that work did not address the structural properties of the events in question, the mechanism identified as responsible—namely, low-infectiousness contagion combined with the occasional encounter with a high-degree node—is largely similar to the one investigated here.

For any node $v_i \in T$ and any subtree $S \in \mathcal{S}$, let $\delta_S(v_i)$ be 1 if $v_i \in S$ and 0 otherwise. Then, summing over all n^2 pairs of nodes, we have

$$\sum_{i,j=1}^{n} \text{dist}(v_i, v_j) = 2n^2 + 2n\sum_{i=1}^{n} \text{depth}(v_i) - 2\sum_{i,j=1}^{n}\sum_{S\in\mathcal{S}} \delta_S(v_i)\delta_S(v_j)$$

$$= 2n^2 + 2n\sum_{i=1}^{n} \text{depth}(v_i) - 2\sum_{S\in\mathcal{S}} |S|^2.$$

The result follows by dividing through by $n(n-1)$ the number of pairs of distinct nodes. □

Theorem 2 shows that $\nu(T)$ can be expressed in terms of the sizes of subtrees of T. Algorithm 1 uses this observation to efficiently compute $\nu(T)$.

Algorithm 1 (Computing $\nu(T)$)
Require: T is a tree rooted at node r

```
function SUBTREE-MOMENTS(T, r)
    if T.size() = 1 then                          ▷ The base case
        size ← 1
        sum-sizes ← 1
        sum-sizes-sqr ← 1
    else                ▷ Recurse over the children of the root r
        for c ∈ r.children() do
            size_c, sum-sizes_c, sum-sizes-sqr_c
                ← SUBTREE-MOMENTS(T, c)
        size ← 0
        sum-sizes ← 0
        sum-sizes-sqr ← 0
        for c ∈ r.children() do
            size ← size + size_c
            sum-sizes ← sum-sizes + sum-sizes_c
            sum-sizes-sqr ← sum-sizes-sqr
                                + sum-sizes-sqr_c
        size ← size + 1
        sum-sizes ← sum-sizes + size
        sum-sizes-sqr ← sum-sizes-sqr + size^2
    return size, sum-sizes, sum-sizes-sqr
function AVERAGE-DISTANCE(T, r)
    size, sum-sizes, sum-sizes-sqr
        ← SUBTREE-MOMENTS(T, r)
    dist_avg ← [2 · size/(size − 1)]×
                [sum-sizes/size − sum-sizes-sqr/size^2]
    return dist_avg
```

Table B.1 **Rank Correlation Between Alternative Measures of Structural Virality**

	Average distance	Relative broadcast	Distinct parent	Average depth
Average distance	1	−0.79	0.73	0.90
Relative broadcast	−0.79	1	−0.98	−0.66
Distinct parent	0.73	−0.98	1	0.61
Average depth	0.90	−0.66	0.61	1

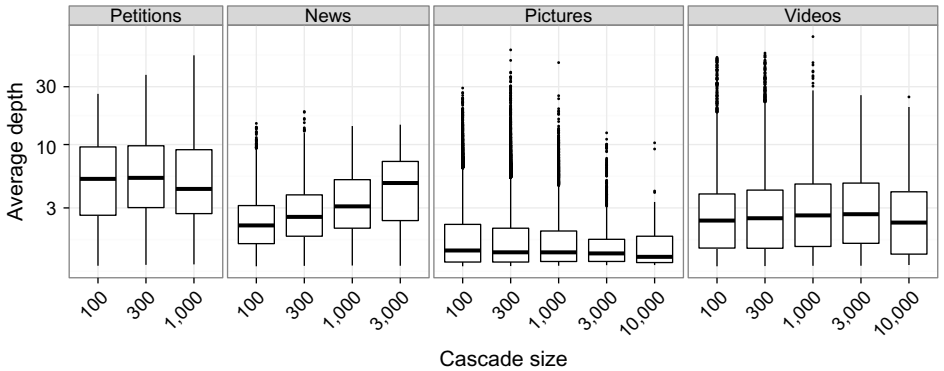
Appendix B. Alternative Measures of Structural Virality

Although we have demonstrated that our particular definition of structural virality is reasonable, there are several other formalizations of the concept that also qualify as reasonable candidates. In particular, here we consider the following three metrics:

1. the relative size of the largest broadcast (i.e., the largest number of children of any single node in the diffusion tree, as a fraction of the total number of nodes in the tree);
2. the probability that two randomly selected nodes have a distinct parent node;
3. the average depth of nodes in the tree.

Simple inspection shows that all three of these alternatives distinguish between the extremes of a single, large broadcast on the one hand and a multigenerational "viral" cascade on the other. However, they all capture subtly different structural aspects of diffusion trees, and also fail for somewhat different pathological cases. Consequently, as with our primary definition above, it is difficult to evaluate the utility of the various metrics on theoretical grounds alone, or even to assess their similarity. In practice, however, we find that they are all highly correlated with our chosen average path length measure, and also with each other. Specifically, Table B.1 shows that when computed over the entire set of empirically observed cascades with at least 100 adopters, $\nu(T)$ has an absolute rank correlation greater than 0.73 with all three alternative measures. Moreover, our empirical results are qualitatively similar regardless of which of these alternative measures of structural virality we apply. For example, Figure B.1 shows the relationship between size and average depth, analogous to Figure 5, and from which essentially the same conclusions could be drawn.

Figure B.1 **Box Plot of an Alternative Measure of Structural Virality—Average Cascade Depth—by Size (on a Log Scale), Separated by Domain**



Note. Lines inside the boxes indicate the medians, whereas the boxes themselves show interquartile ranges.

Thus, although we cannot rule out the possibility that a superior metric to ours can be defined, we can at least substantiate two related claims: first, that our choice of metric is at least roughly as good as a number of other plausible candidates, and second, that our substantive findings are robust with respect to the particular manner in which we formalize the concept of structural virality.

Appendix C. Tree Construction Method

Here we describe the process of constructing a diffusion tree for a particular piece of content (e.g., a given URL). Trees are composed of one node for each user who has adopted the content, and each edge links a user back to an inferred "parent." After each adoption has been identified as either a root or the child of another post, we construct the cascade of adoptions.

In an ideal setting we would have access to this information for each adoption, but in practice these details are not always available. The best-case scenario is use of Twitter's official retweet functionality, which enables a user to effectively forward a tweet that was originally authored by someone else. Attribution is clear in these cases, and tree construction would be relatively straightforward if all adoptions were of this form. Unfortunately, however, users also repost content using a variety of unofficial conventions, which complicate the attribution task. For instance, the unofficial retweet convention amounts to copying the text of a tweet and prepending "RT @username" to credit another individual. Twitter treats these posts as originally authored content and has no formal way of linking them back to original posts. Finally, users may forego crediting a source entirely, in which case one must make an educated guess about who (if anyone) in their feed exposed them to the content and who should be credited as responsible for their adoption.

We decompose the process of inferring a parent into two steps, described in detail below. We estimate that our inference procedure correctly identifies the parent of an adoption in approximately 95% of instances.

1. *Identify potential parents.* For each user who adopts a piece of content, we identify a set of "potential parents," defined as individuals whose adoption of a piece of content appears in the focal user's timeline prior to the focal user's adoption. In other words, potential parents are the set of individuals who are likely to have exposed the user to the adopted content. To identify these potential parents, we note that a user's timeline contains (1) all posts originally authored by the user's friends and (2) tweets authored by others that at least one of the user's friends has "officially retweeted" using Twitter's built-in reposting functionality. In particular, any tweet appears at most once in a user's timeline regardless of how many of his or her friends have officially retweeted it.[19] To compute the set of potential parents for a given adoption, we join activity from the Twitter Firehose application programming interface (API), which provides details about each tweet, with the Twitter follower graph, which provides the listing of who follows whom.

[19] Any nonofficial reposting—e.g., using the "RT @username" convention—is considered originally authored, resulting in potentially repeated content in a user's timeline.

2. *Infer a single parent.* We now identify the single most likely parent from the set of all potential parents of a given adoption. To do this, we consider three cases based on how the focal user posted the content.

a. *Official retweet.* If the focal user officially retweeted a post that appeared in their timeline (i.e., retweeted the post via Twitter's built-in functionality), then the Twitter API provides the ID of the original tweet. We then use this information to identify the individual who introduced the post to the user's timeline as the parent. We note that the parent need not be the original author of the tweet—for example, in the case of a friend who retweeted a third party, as described above. Also, users occasionally officially retweet content that did not appear in their timelines (e.g., because they discovered it by browsing); in these cases we treat the focal user as a "root" and do not assign a parent. Overall, in these official retweet cases—which constitute 65% of the instances we consider—we almost certainly correctly attribute the tweet.

b. *Accredited repost.* In the case of a nonofficial retweet, credit may still be present in the form of a mentioned user, for example, using the "RT @username" convention. We identify as the parent the individual who most recently introduced a post of that content, authored by the mentioned user, to the focal user's timeline. This mentioned user may be a friend of the focal user, in which case the friend is assigned as the parent. Alternatively, the mentioned user may be a third party—e.g., a friend of a friend. In this case, the friend who most recently mentioned the accredited user along side the piece of content is identified as the parent. As above, if no such friend can be identified, we treat the focal user as a root and do not assign a parent. Accredited posts constitute 10% of the adoptions we analyze, and as in the case of official retweets, the inferred parent is almost certainly correct.

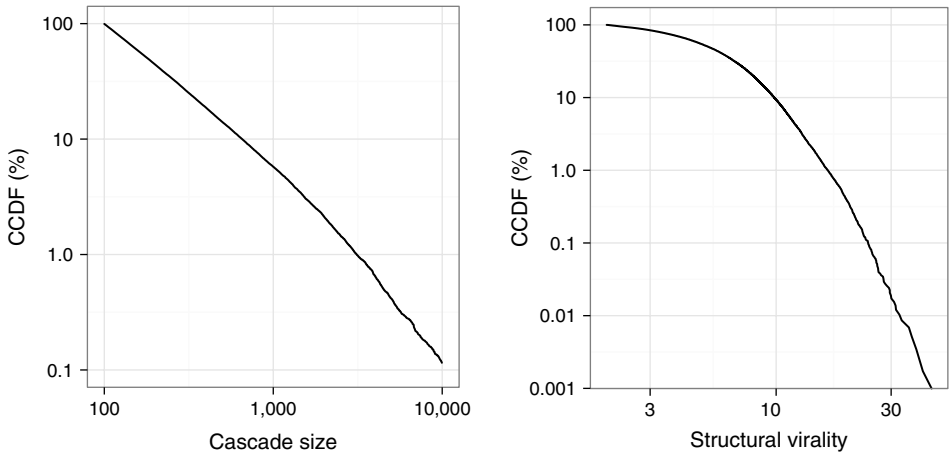
c. *Uncredited repost.* In this final, case we lack any explicit information about how the user was exposed to the content and simply assign as the parent the friend who most recently introduced the content to the focal user's timeline. If no such friend exists, we again treat the focal user as a root. To assess the accuracy of our inference strategy in this case, we apply it to the set of official retweets, for which we are fairly certain which individual is the parent of any given adoption. We find that the most-recent-introduction heuristic correctly identifies the parent 79% of the time.

Since our inference procedure almost certainly identifies the correct parent in the first two cases—official retweets and accredited reposts, which together account for 75% of adoptions—and since we estimate 79% accuracy for the remaining 25% of adoptions, we conclude that the overall accuracy of our parent inference strategy is 95%.

Appendix D. Off-Channel Diffusion

Although our empirical findings are qualitatively quite similar across the four distinct domains studied above, it is possible that all four suffer from one of two systematic biases that might affect our conclusions. First, a potential problem with studying the diffusion of external content on Twitter (e.g., news stories from the *New York Times* and videos from YouTube) is that the same content may also spread via other channels, such as Facebook or email. As a result of this "off-channel" diffusion, two individuals on Twitter who appear

Figure D.1 Size and Structural Virality Distributions on a Log–Log Scale for Popular Hashtag Cascades, Containing at Least 100 Adopters

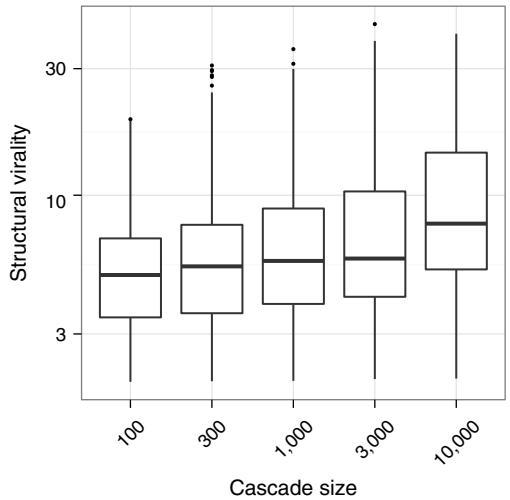


Note. CCDF, complementary cumulative distribution function.

to have introduced the same piece of content independently may in fact be connected, thus leading us to mistakenly treat a single diffusion tree as two disjoint events. A second concern is that our use of reposting rather than retweeting also potentially biases our data. Specifically, user–follower similarity (i.e., homophily) may lead connected users to post the same content independently in close temporal sequence, leading us to conflate similarity with influence (Shalizi and Thomas 2011, Aral et al. 2009, Lyons 2011).

To check that off-channel diffusion does not systematically bias our findings, we consider the diffusion of Twitter-specific "hashtags"—short fragments of text used to indicate the topic of a tweet. Because such hashtags are less likely to have originated outside of Twitter, and because for the same reason they are less likely to migrate off of Twitter, these data are correspondingly less susceptible to any biases associated with off-channel diffusion. Moreover, to ensure as much as possible that we are considering only on-Twitter uses of hashtags, we restrict our sample to "long" hashtags, which are especially unlikely to be used elsewhere. To define "long," we note that hashtags on Twitter are generally written in camel case (e.g., #CamelCase). Treating each substring that begins with a capitalized letter and ends immediately before the next capitalized letter as a "word," we trace the diffusion of hashtags that include five or more such words (e.g., #ThisIsALongHashtag). As infrequent as these long hashtags are relative to hashtags in general, they are still plentiful, amounting to 58,000 cascades with at least 100 adopters. Figures D.1 and D.2 show that the diffusion of these long hashtags yields qualitatively similar results to our primary analysis, suggesting that off-channel diffusion is not driving our findings.

Figure D.2 Box Plot of Structural Virality by Size on a Log–Log Scale for Hashtag Cascades



Note. Lines inside the boxes indicate the median structural virality, whereas the boxes themselves show interquartile ranges.

References

Adar E, Adamic LA (2005) Tracking information epidemics in blogspace. *IEEE/WIC/ACM Internat. Conf. Web Intelligence* (Institute of Electrical and Electronics Engineers, Piscataway, NJ).

Anderson RM, May RM (1991) *Infectious Diseases of Humans* (Oxford University Press, Oxford, UK).

Aral S, Muchnik L, Sundararajan A (2009) Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proc. Natl. Acad. Sci. USA* 106(51):21544–21549.

Bakshy E, Karrer B, Adamic LA (2009) Social influence and the diffusion of user-created content. *Proc. Tenth ACM Conf. Electronic Commerce* (Association for Computing Machinery, New York), 325–334.

Bakshy E, Hofman JM, Mason WA, Watts DJ (2011) Everyone's an influencer: Quantifying influence on twitter. *Proc. Fourth ACM Internat. Conf. Web Search and Data Mining* (Association for Computing Machinery, New York), 65–74.

Barabási AL, Albert R (1999) Emergence of scaling in random networks. *Science* 286(5439):509–512.

Bass FM (1969) A new product growth for model consumer durables. *Management Sci.* 15(5):215–227.

Bass FM (2004) Comments on "a new product growth for model consumer durables the bass model." *Management Sci.* 50(12 supplement):1833–1840.

Centola D (2010) The spread of behavior in an online social network experiment. *Science* 329(5996):1194–1197.

Chierichetti F, Kleinberg J, Liben-Nowell D (2011) Reconstructing patterns of information diffusion from incomplete observations. *Adv. Neural Inform. Processing Systems*, Vol. 24 (Neural Information Processing Systems Foundation, La Jolla, CA).

Clauset A, Shalizi CR, Newman MEJ (2009) Power-law distributions in empirical data. *SIAM Rev.* 51(4):661–703.

Coleman J, Katz E, Menzel H (1957) The diffusion of an innovation among physicians. *Sociometry* 20(4):253–270.

Dodds PS, Watts DJ (2004) Universal behavior in a generalized model of contagion. *Phys. Rev. Lett.* 92(21):218701.

Dow PA, Adamic LA, Friggeri A (2013) The anatomy of large Facebook cascades. *Proc. Seventh Internat. AAAI Conf. Weblogs Soc. Media* (AAAI Press, Palo Alto, CA).

Fichman RG (1992) Information technology diffusion: A review of empirical research. *Proc. 13th Internat. Conf. Inform. Systems* (University of Minnesota, Minneapolis), 195–206.

Goel S, Watts DJ, Goldstein DG (2012) The structure of online diffusion networks. *Proc. 13th ACM Conf. Electronic Commerce* (Association for Computing Machinery, New York), 623–638.

Golub B, Jackson MO (2010) Using selection bias to explain the observed structure of Internet diffusions. *Proc. Natl. Acad. Sci. USA* 107(24):10833–10836.

Granovetter M (1978) Threshold models of collective behavior. *Amer. J. Sociol.* 83(6):1420–1443.

Hoang T-A, Lim E-P (2012) Virality and susceptibility in information diffusions. *Proc. Sixth Internat. AAAI Conf. Weblogs Soc. Media* (AAAI Press, Palo Alto, CA).

Ijiri Y, Simon HA, Bonini CP, van Wormer TA (1977) *Skew Distributions and the Sizes of Business Firms* (North-Holland Publishing Company, New York).

Iyengar R, Van den Bulte C, Valente TW (2010) Opinion leadership and social contagion in new product diffusion. *Marketing Sci.* 30(2):195–212.

Jenders M, Kasneci G, Naumann F (2013) Analyzing and predicting viral tweets. *Proc. 22nd Internat. Conf. World Wide Web Companion* (International World Wide Web Conferences Steering Committee, New York), 657–664.

Kempe D, Kleinberg J, Tardos E (2003) Maximizing the spread of influence through a social network. *Proc. Ninth ACM SIGKDD Internat. Conf. Knowledge Discovery and Data Mining* (Association for Computing Machinery, New York), 137–146.

Kermack WO, McKendrick AG (1927) A contribution to the mathematical theory of epidemics. *Proc. Roy. Soc. Lond. A* 115(772): 700–721.

Kupavskii A, Ostroumova L, Umnov A, Usachev S, Serdyukov P, Gusev G, Kustarev A (2012) Prediction of retweet cascade size over time. *Proc. 21st ACM Internat. Conf. Inform. Knowledge Management* (Association for Computing Machinery, New York), 2335–2338.

Kwak H, Lee C, Park H, Moon S (2010) What is twitter, a social network or a news media? *WWW '10: Proc. 19th Internat. Conf. World Wide Web* (Association for Computing Machinery, New York), 591–600.

Leskovec J, Adamic LA, Huberman BA (2007) The dynamics of viral marketing. *ACM Trans. Web* 1(1):5.

Leskovec J, Singh A, Kleinberg J (2006) Patterns of influence in a recommendation network. *Adv. Knowledge Discovery Data Mining* 3918:380–389.

Liben-Nowell D, Kleinberg J (2008) Tracing information flow on a global scale using Internet chain-letter data. *Proc. Natl. Acad. Sci. USA* 105(12):4633–4638.

Lloyd AL, May RM (2001) How viruses spread among computers and people. *Science* 292(5520):1316–1317.

Lopez-Pintado D, Watts DJ (2008) Social influence, binary decisions and collective dynamics. *Rationality Soc.* 20(4):399–443.

Lyons R (2000) Phase transitions on nonamenable graphs. *J. Math. Phys.* 41(3):1099–1126.

Lyons R (2011) The spread of evidence-poor medicine via flawed social-network analysis. *Statist., Politics, Policy* 2(1).

Ma Z, Sun A, Cong G (2013) On predicting the popularity of newly emerging hashtags in Twitter. *J. Amer. Soc. Inform. Sci. Tech.* 64(7):1399–1410.

Mahajan V, Peterson RA (1985) *Models for Innovation Diffusion*, Vol. 48 (Sage, Newbury Park, CA).

Mohar B, Pisanski T (1988) How to compute the wiener index of a graph. *J. Math. Chemistry* 2(3):267–277.

Newman MEJ (2005) Power laws, Pareto distributions and Zipf's law. *Contemporary Phys.* 46(5):323–351.

Pastor-Satorras R, Vespignani A (2001) Epidemic spreading in scale-free networks. *Phys. Rev. Lett.* 86(14):3200–3203.

Petrovic S, Osborne M, Lavrenko V (2011) RT to win! Predicting message propagation in Twitter. *Proc. Fifth Internat. AAAI Conf. Weblogs Soc. Media* (AAAI Press, Palo Alto, CA).

Rogers EM (1962) *Diffusion of Innovations* (Free Press, New York).

Shalizi CR, Thomas AC (2011) Homophily and contagion are generically confounded in observational social network studies. *Sociol. Methods Res.* 40(2):211–239.

Sun E, Rosenn I, Marlow C, Lento T (2009) Gesundheit! Modeling contagion through facebook news feed. *Proc. Third Internat. AAAI Conf. Weblogs Soc. Media* (AAAI Press, Palo Alto, CA).

Toole JL, Cha M, González MC (2012) Modeling the adoption of innovations in the presence of geographic and media influences. *PLoS One* 7(1):e29528.

Tsur O, Rappoport A (2012) What's in a hashtag? Content based prediction of the spread of ideas in microblogging communities. *Proc. Fifth ACM Internat. Conf. Web Search and Data Mining* (Association for Computing Machinery, New York), 643–652.

Valente TW (1995) *Network Models of the Diffusion of Innovations*, Quantitative Methods in Communication Series (Hampton Press, Cresskill, NJ).

Van den Bulte C, Lilien GL (2001) Medical innovation revisited: Social contagion versus marketing effort1. *Amer. J. Sociol.* 106(5):1409–1435.

Walther JB, Carr CT, Choi SSW, DeAndrea DC, Kim J, Tong ST, Van Der Heide B (2010) Interaction of interpersonal, peer, and media influence sources online. *A Networked Self: Identity, Community, and Culture on Social Network Sites*, Vol. 17 (Routledge, London).

Watts DJ (2002) A simple model of information cascades on random networks. *Proc. Natl. Acad. Sci. USA* 99(9):5766–5771.

Wiener H (1947) Structural determination of paraffin boiling points. *J. Amer. Chemical Soc.* 69(1):17–20.

Wu S, Hofman JM, Mason WA, Watts DJ (2011) Who says what to whom on Twitter. *Proc. 20th Internat. Conf. World Wide Web* (Association for Computing Machinery, New York), 705–714.

Yang J, Counts S (2010) Predicting the speed, scale, and range of information diffusion in Twitter. *Proc. Fourth Internat. AAAI Conf. Weblogs Soc. Media* (AAAI Press, Palo Alto, CA), 355–358.

Yang J, Leskovec J (2010) Modeling information diffusion in implicit networks. *Proc. 10th IEEE Internat. Conf. Data Mining* (IEEE Computer Society, Washington, DC), 599–608.

Young PH (2009) Innovation diffusion in heterogeneous populations: Contagion, social influence, and social learning. *Amer. Econom. Rev.* 99(5):1899–1924.