The structural virality of online diffusion

Sharad Goel[1], Ashton Anderson[2], Jake Hofman[1], and Duncan Watts[1]

[1]Microsoft Research
[2]Stanford University

DRAFT

Abstract

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 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 dataset of a billion diffusion events on Twitter, including the propagation of news stories, videos, images, and petitions. We find that the very largest observed events nearly always exhibit high structural virality, providing some of the first direct evidence that many of the most popular products and ideas grow through person-to-person diffusion. However, medium-sized events—having thousands of adopters—exhibit surprising structural diversity, and regularly grow via both broadcast and viral mechanisms. We find that these empirical results are largely consistent with a simple contagion model characterized by a low infection rate spreading on a scale-free network, reminiscent of previous work on the long-term persistence of computer viruses.

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)

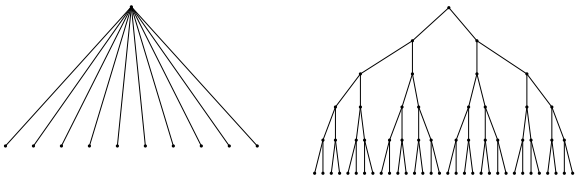

Figure 1: A schematic depiction of broadcast versus viral diffusion, where nodes represent individual adoptions and edges indicate who adopted from whom.

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 page of the most popular news websites attract a similar number of daily visitors—hence the mere observation that something is popular, or even that it became popular rapidly, is not sufficient to establish that it spread in a manner that resembles social contagion.

Fig. 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 multi-generational branching process in which any one node directly "infects" only a few others. Although the stylized patterns in Fig. 1 are intuitively plausible and also easily distinguishable from one another, real diffusion events are unlikely to conform precisely to either ideal type. 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.

Differentiating systemically between broadcast vs. viral diffusion, or indeed any combination of the two, 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 had not been previously formulated 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

[1] Even models of social contagion that do not correspond precisely to the mechanics of biological infectious disease (e.g., "threshold models" (Granovetter 1978) make different assumptions regarding the non-independence 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)

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). To understand the underlying structure of the event, and in particular to disambiguate between the distinct possibilities of viral and broadcast diffusion, 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), e-commerce (Leskovec et al. 2006), multiplayer gaming (Bakshy et al. 2009), and social networking (Sun et al. 2009, Bakshy et al. 2011, Goel et al. 2012).

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 in order 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, while "viral hits" appear at a rate closer to one in a million. Consequently, in order 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 formalizes the intuitive distinction between broadcast and viral diffusion, and that interpolates between these two extremes in a way that allows us to empirically disambiguate between different structures. We emphasize that our structural approach to virality is a complement to, not a 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). That is, whereas 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 the properties of the network over which the diffusion occurs (or, more generally, of the contact process), or the timescales associated with adoption—our measure of structural virality is concerned exclusively with characterizing resulting adoption patterns. In particular, structural virality is designed to capture the diversity and complexity of diffusion structures that arise in very large event populations in a way that is independent of any specific generative model of contagion.

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 rigorously estimate the frequency of viral cascades, quantify the diversity in the structure of cascades, and investigate the relationship between cascade size and structure. A priori, it is not obvious how empirical cascades are structured. According to one tradition of theoretical models (Coleman et al. 1957, Toole et al. 2012) viral transmission can only take place once a critical threshold value of infectiousness is exceeded. If this intuition is correct, one should expect large events either to be broadcasts (i.e., when the critical threshold is not exceeded) or to be multigenerational cascades (when it is), but not combinations of the two. Alternative theories (Pastor-Satorras and Vespignani 2001, Lloyd and May 2001), however, lack such a critical threshold, allowing for a smooth, continuous transition between broadcasts and viral cascades. Regarding the relationship between size and structure, it could be the case that the most popular content is also extremely viral; but equally it could be that such successful products are the direct effect of mass media (i.e., a single large broadcast); or it could be that some combination of broadcasts and word-of-mouth is common for large events. Depending on the relative importance of broadcast vs. 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 enables us to distinguish between such possibilities, wherein we find a surprisingly wide array of diverse cascade structures, and low correlation between popularity and virality.

The third and final contribution of this paper is to compare our empirical observations of cascade structure to predictions from a simple generative model of diffusion. Specifically, we conduct large-scale simulations of a simple disease-like contagion model, similar to the original Bass model of product adoption (Bass 1969), on a network designed to capture the main features of the Twitter follower graph. The large scale is important in two respects: first, because large diffusion events are so rare, we must conduct on the order of one billion simulations per parameter setting, necessitating over 100 billion simulations in total; and second, to adequately imitate the topology of the empirical network, where the number of followers per user spans several orders of magnitude (roughly, from tens to tens of millions), each simulation must be conducted on a network on the order of 100 million nodes. From a purely computational standpoint, therefore, attempting to replicate empirical patterns concerning rare events in large-scale networks is a challenging simulation task. Nevertheless, we find that when simulated on the appropriate scale, a relatively simple model can capture many of the features of our empirical data.

2. Defining Structural Virality

We now turn to our first goal of formally quantifying the structural virality of diffusion trees, and in particular, of disambiguating between the broadcast and multigenerational, viral distribution channels depicted in Fig. 1. A natural choice for such a metric is 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 a 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 non-representative 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 Fig. 1), where $\nu(T) \approx 2$. Moreover, a complete k-ary tree (as in Fig 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 $\nu(T)$ avoids the problem cases associated with depth and average depth, it is possible to construct pathological 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 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 further below.[4]

[2]Naive computation of $\nu(T)$ requires $O(n^2)$ time; however, as discussed in Appendix 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, while 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]In addition, in Appendix B we examine three additional measures of structural virality, including average depth. We find that in practice these metrics are highly correlated with average distance $\nu(T)$, and that our results are robust to the specific formalization we use.

3. Data and Methods

Our primary analysis is based on approximately one 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.[5] 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.[6] In total, we observe roughly 750 million unique pieces of content; however, because individual pieces of content can be posted by multiple users, we observe approximately 1.4 billion "adoptions" (i.e., posting of content).[7] Although our dataset is not a complete census of web content shared on Twitter, it does include the vast majority and hence is essentially unbiased at least with respect to tweets linking to web content.[8] 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 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.

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.[9] 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

[5]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. To check that our results aren't driven by homophily, in Appendix Appendix C. we repeat the analysis on cascades constructed from official retweets only and find that our results are qualitatively unchanged.

[6]For news: 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, msnbc.msn.com. For video: youtube.com, m.youtube.com, youtu.be, vimeo.com, livestream.com, twitcam.livestream.com, ustream.tv, twitvid.com, mtv.com, vh1.com. For images: twitpic.com, instagr.am, instagram.com, yfrog.com, p.twimg.com, twimg.com, i.imgur.com, imgur.com, img.ly, flickr.com. For petitions: change.org, twitition.com, kickstarter.com.

[7]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.

[8]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. In Appendix C, however, we conduct a separate analysis of tweets containing hashtags (instead of links to web content), and find qualitatively similar results.

[9]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.

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 if 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 is labeled the focal adopter's "parent".[10] In this way we construct disjoint diffusion trees for each independent introduction of a piece of content, one for each root, where we note that the same piece of content—say a particular news story or YouTube video—can be introduced independently many times, and hence generate many distinct diffusion trees of possibly widely varying sizes, where the collection of all such trees associated with single unique piece of content constitutes a "forest" for that piece.

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 1.4—meaning that for every ten introductions of content, there are on average four 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 datasets. Exploiting the fact that we have a much larger dataset 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 0.03% of diffusion trees containing at least 100 nodes (Fig. 2), a requirement that leaves us with roughly 1 out of every 3,000 cascades, and thus reduces the number of cascades we study in detail from approximately 1 billion to 343,000.

4. Empirical Results

We begin our empirical analysis by examining a stratified random sample of empirically observed cascades, ordered by their structural virality $\nu(T)$ (Fig. 3). Specifically, cascades with between 100 and 1000 adopters were ranked by $\nu(T)$ and logarithmically binned; a random cascade was then drawn from each bin, where we note that this exercise was performed only once to avoid hand selection of the best "random" sample. We make two key observations about the sampling of cascades in Fig. 3.

First, 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

[10] In the case of multiple friends adopting before the focal adopter, it is clearly possible to make other choices of parent—for example, the first friend to adopt the content, or a random choice. Given the speed with which content disappears from view on a typical Twitter feed, however, selecting the friend who most recently adopted before the focal adopter appears to be a reasonable choice; that is, when a user reposts a piece of content, the most recently posted instance is likely to be the proximate cause.

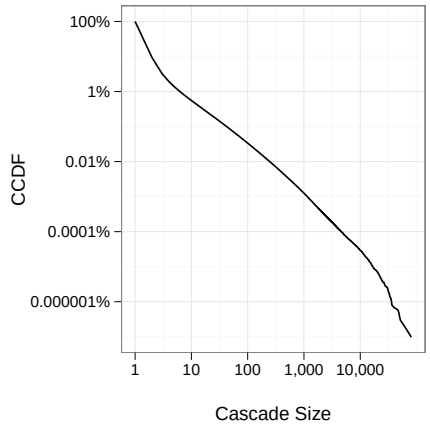

Figure 2: Distribution of cascade sizes on a log-log scale, aggregated across the four domains we study: videos, news, pictures, and petitions.

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. Moreover, as shown in the cumulative adoption curves below each cascade in Fig. 3, all the selected diffusion events experience a phase of rapid growth before leveling off. Although these adoption curves are not identical, our measure of structural virality does seem to more effectively and directly quantify differences in the underlying cascade structures than does this traditional signal of virality based on aggregate data.

Second, although the structures in Fig. 3 are all of similar size (i.e., have similar aggregate numbers of adopters), they exhibit remarkable diversity in structure, from a pure broadcast ($\nu(T) = 2$, top left) to an ideal-type branching structure ($\nu(T) = 49$, bottom right), with numerous intermediate variations in between. As noted above, the classical literature on diffusion often posits a critical threshold for virality, suggesting a sharp break between cascades that are viral and those that are not. If this intuition is correct, one would expect that relatively large diffusion events such as those captured in the Fig. 3 would be either pure broadcasts or viral cascades but not intermediate forms. More generally, even if one regarded the pure critical threshold model as overly simplistic, one might still expect only a handful of canonical forms to account for the majority of large events: for example, some events spread exclusively via broadcast, while others spread exclusively via word-of-mouth, and others still spread by some particular "typical" combination of the two. Fig. 3, however, shows that in fact there is no typical pattern at all—rather, along with examples at both extremes of the broadcast-to-viral spectrum, we see an almost continuous distribution of fine-grained variations in-between.

4.1. Examining Popularity and Structural Virality

Fig. 3 shows that structural virality can vary widely for events of the same size, or popularity. By keeping size fixed, however, it does not allow us to conclude anything about the relative frequency

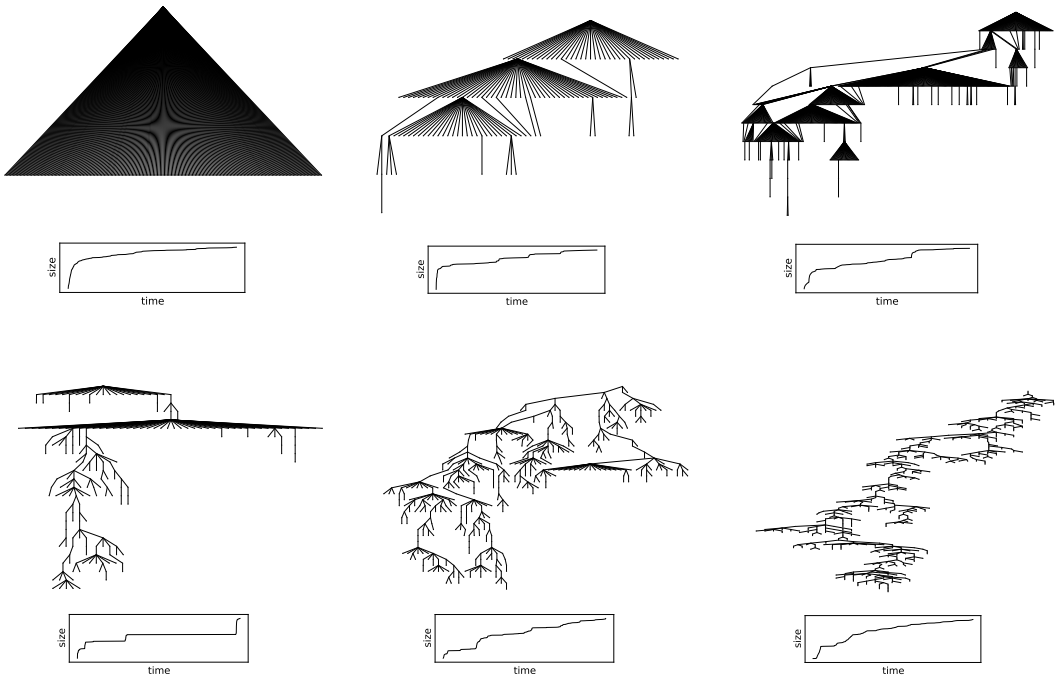

Figure 3: A random sample of cascades stratified and ordered by increasing structural virality, ranging from 2 to 50. For ease of visualization, cascades were restricted to having between 100 and 1000 adopters. Cumulative adoption curves (i.e., total cascade size over time) are shown below each cascade.

of structural virality for events of a given size, nor does it reveal how these frequencies change with increasing popularity. With Fig. 3 as motivation, therefore, we now apply $\nu(T)$ first to investigate the structure of online diffusion more systematically across our four content domains—news, videos, images, and petitions—and second to examine the relationship between structural virality and popularity.

To begin with, Fig. 4 shows both the size distribution of cascades larger than 100 adopters for all four domains (Fig. 4A), and also the corresponding distributions of structural virality (Fig. 4B). In spite of several qualitative domain differences, cascades with high structural virality regularly occur in all four domains. For example, about 3% of large news cascades (i.e., cascades with at least 100 adopters) have structural virality of 10 or more, as do about 10% of large video and image cascades. Strikingly, about 50% 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. Though high structural virality is relatively common among these large cascades, we recall that only a small minority of cascades—roughly 1 in 3000, across domains—attracts at least 100 adopters. While cascades that are both large and viral are therefore rare, we still observe hundreds of such events in our data.[11] As noted earlier, though there is a large theoretical literature on modeling viral events, there has in fact been scant direct observation of their occurrence (Liben-Nowell and Kleinberg 2008, Dow et al. 2013), in part because the ability to track diffusion cascades is relatively new, and in part because viral events are exceedingly rare. The observation from Fig. 4 that cascades with high structural virality do in fact occur with regularity and in a variety of domains, thus provides some of the first direct evidence of viral information propagation.

Next we observed that across all four domains we see wide variation in both size, which ranges from 100 (our self-imposed popularity threshold) up to several tens of thousands, and also structural virality, which ranges from 2 to over 100 (meaning that adopters in these cascades are on average as few as 2 and as many as 100 steps apart), thus confirming our earlier observation about the diverse and complex nature of information diffusion, illustrated visually in Fig. 3. In particular, in contrast to classical 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.

Finally, while the four domains exhibit qualitatively similar size and virality distributions, Fig. 4 also reveals some differences between domains. In particular, Fig. 4B suggests an ordering of domains by virality, with petitions the most viral, followed by videos and images together, followed by news stories. Popular petitions, that is, are much more likely to have grown virally than popular news stories, images or videos: whereas nearly 50% of popular petitions have structural virality of at least 10, less than 5% of news stories do.

[11] The cascade in the bottom right corner of Fig. 3 is one such example, comprising 618 nodes and having spread for 132 generations.

4.2. Relationship between Popularity and Structural Virality

By grouping together all cascades with at least 100 adopters, Fig. 4 on its own says little about the relationship between size and virality. As pointed out earlier, that is, depending on the empirically observed preponderance of broadcasts in small versus large events, the relationship between size and structural virality could be positive (larger 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 (e.g., all events, large and small, are dominated by broadcasts, where the variation in size is determined by the largest such broadcast). 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, however, 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.

Fig. 5 investigates this question, showing the distribution of structural virality conditional on cascade size for each domain. First, Fig. 5 shows that across all four domains, average structural virality does increase with cascade size. For example, video cascades with approximately 100 adopters have median virality 5, compared to 20 for cascades of size 10,000. In fact, the largest cascades we observe—video cascades with 30,000 adoptions—are nearly always viral, with 89% of such events having structural virality greater than 10. We emphasize that such a result need not have been the case. For example, if the most popular videos were typically the product of a single, large broadcast, we would expect structural virality to be relatively low—recall that a pure broadcast has structural virality approximately equal to two. Instead, we find that though small cascades are dominated by broadcast-like diffusion, the very largest events are likely to have grown virally.[12]

While the very largest cascades are almost always viral, Fig. 5 also shows that cascades with high structural virality occur—and often quite frequently—at nearly every size. For example, while the median virality of news stories with approximately 1,000 adoptions is 5, the interquartile range includes values up to 8 and extremes as high as 35. Moreover, the converse also applies: except for the very largest cascades, we find examples of broadcasts and near-broadcasts (i.e., with $\nu(T) \approx 2$) at all sizes and across all domains. Put another way, whereas past theoretical and empirical work has focused almost exclusively on a handful of the very largest events, our dominant finding is that there is extraordinary diversity at almost every scale.

There are thus two ways to view the relationship between popularity and structure: first, larger cascades are more likely to grow virally than smaller ones; but second, size alone is a poor predictor of structure, in part due to the significant variation in structural virality that we observe at nearly all sizes. In quantitative terms, among cascades having at least 100 adopters, the correlation between size and structural virality is 0.36, indicating a positive but noisy relationship between the two measures.[13] Put another way, although size and structural virality are positively related, knowing

[12] We 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 Appendix B.).

[13] Clearly, the precise numerical value of this correlation depends on size range over which it is computed. For example, among cascades having between 100 and 1000 adopters, the correlation between size and structural virality is 0.26.

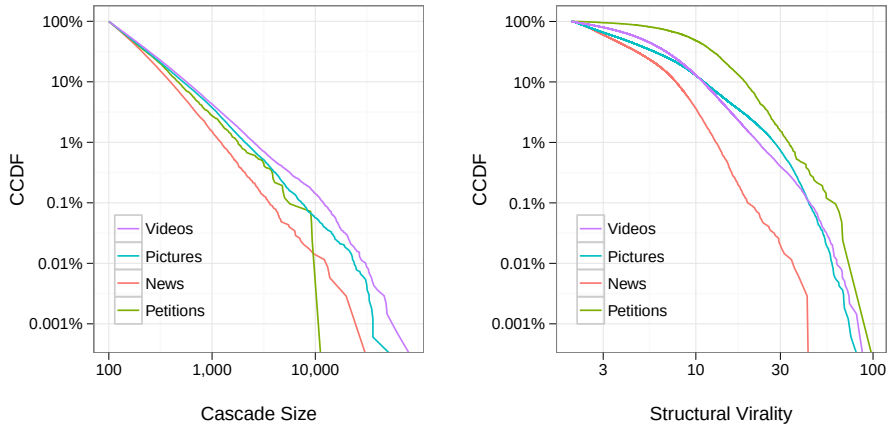

Figure 4: Size and structural virality distributions on a log-log scale for cascades containing at least 100 adopters, separated by domain.

only the popularity of a cascade it remains difficult to determine whether it grew via broadcast or viral means.

Finally, with respect to domain differences, Fig. 5 shows that news, images, and videos all follow a similar pattern. That is, although we see larger events for images than for news, and larger events still for videos, the corresponding size bins in all three panels show a remarkably similar distribution of structural virality. Controlling for size, in other words, news, images and videos are similar with respect to structural virality. Making the same comparison with petitions, however, it is equally striking that popular petitions are clearly more viral at each size scale than any of the other three domains. Although we can only speculate about what drives such differences, they may in part be due to a relative dearth of large broadcast channels for petitions, as well as possibly different incentives for sharing such content.

5. Theoretical Modeling

Our empirical observations raise an interesting theoretical problem. On the one hand, the vast majority of diffusion events are quite small, and accordingly lack much structure (Goel et al. 2012). On the other hand, those events that do become large exhibit striking structural diversity; moreover, the size of these cascades is only moderately related to their structural virality. Although it may be tempting to treat these small and large events as products of distinct propagation mechanisms, we find, perhaps surprisingly, that a simple model of diffusion captures both of these empirically observed aspects of online diffusion.

Specifically, we study 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, initially to the specific context of new product adoption, where it is known as the "Bass model" (Bass 1969), and subsequently to a wide range of other contexts including the propagation of links over a network of blogs (Leskovec et al. 2007). Reflecting its origins in math-

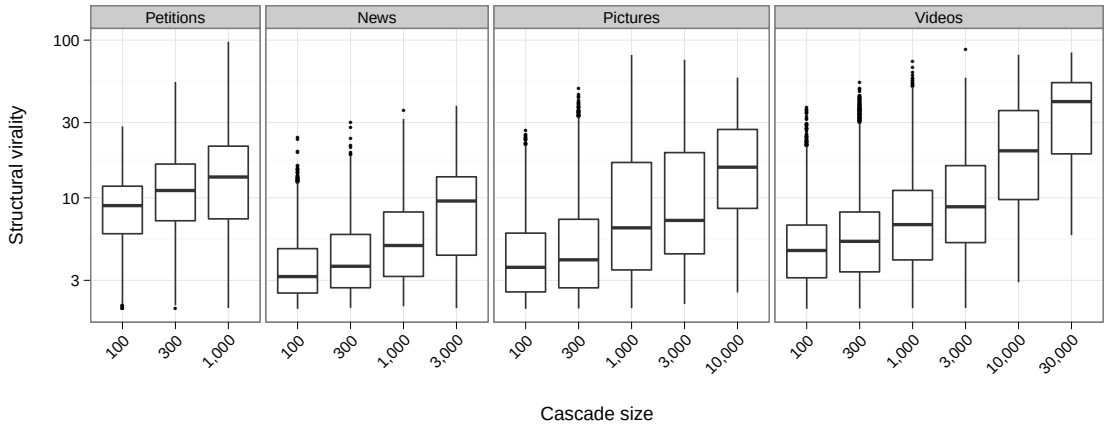

Figure 5: Boxplot of structural virality by size on a log-log scale, separated by domain. Lines inside the boxes indicate median structural virality, while the boxes themselves show interquartile ranges.

ematical epidemiology, the model is named for the three states—"susceptible," "infectious," and "recovered"—that each node in the network can occupy.[14] Specifically, 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. After one time step, infected nodes are removed from the dynamics, meaning that they can no longer infect others, nor become reinfected. As the degree distribution on Twitter is heavy tailed (Bakshy et al. 2011), and as the empirically observed cascades exhibit periodic bursts of adoption attributable to high degree individuals, we simulate the model on a scale-free random network (Barabási and Albert 1999). The process is therefore governed by just two parameters: (1) the exponent α describing the power-law degree distribution of the network, with lower values of α resulting in a heavier tail; and (2) the intrinsic "infectiousness" β of the content spreading over the network.

Before proceeding, it is helpful to introduce the quantity $R_0 = k\beta$, known in mathematical epidemiology as the "basic reproduction number" of a disease, where k is the average node degree of the network (i.e., the number of opportunities a node typically has to infect others). As alluded to earlier, a standard result is that in networks with a degree distribution having finite variance, the condition $R_0 = 1$ constitutes a critical threshold separating two regimes: the "subcritical" regime $R_0 < 1$, in which all novel infections die out before infecting more than an infinitesimal fraction of an infinite network; and the "supercritical" regime $R_0 > 1$ in which the contagion spreads exponentially fast. In scale-free networks, it has been shown that no such critical threshold exists; that is, even for $R_0 < 1$, a non-zero—but potentially very small—fraction of an infinite network can be infected in equilibrium (Pastor-Satorras and Vespignani 2001, Lloyd and May 2001).[15] Nevertheless, we find

[14]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.

[15]Technically, the expression for R_0 is: $R_0' = k\beta\left(1 + (\sigma/k)^2\right)$, where σ^2 accounts for the variance of the degree distribution (Lloyd and May 2001). Clearly, when σ is infinite, as is the case in infinite, scale-free networks, then $R_0' > 1$ for any finite β, hence no critical threshold exits.

it helpful to retain the sub/super critical distinction, as it is implicit in much of the discussion of "going viral."

Simulation details. Each realization of the simulation commences with an entirely susceptible population within which a single individual is randomly chosen to be the initially infected "seed," and proceeds until no further infections can take place. We simulated the SIR model on a scale-free random network comprising 25 million nodes constructed using the configuration method (Newman 2005, Clauset et al. 2009): for each node in the network its degree was first randomly selected according to a discrete power law 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. Sweeping over the two parameters, α and β, we simulated content of varying infectiousness diffusing over networks with varying degree skew.

From our data, we know that large events are extremely rare, hence we focus our attention on relatively small values of β and simulate a correspondingly large number of diffusion events. In particular, setting $R_0 = \beta\bar{d}$ where $\bar{d}$ is the average network degree, we restrict to the region $0 < R_0 < 1$. Fig. 6 shows the results of nearly 100 billion simulations, with one billion cascades generated for each parameter setting (α, β), roughly congruent with the number of empirical observations.

Most importantly, Fig. 6 shows that for certain parameters—$r \approx 0.7$ and $\alpha \approx 2.3$—the model recapitulates several important features of our empirical data. First, Fig. 6A 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, Fig. 6B shows that mean virality for these parameters is 5, which again is in line with our observations. Third, Fig. 6C shows that the correlation between size and structural virality is also in the observed range. Finally, Fig. 7 shows the distribution of virality conditional on 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) is considerably less in Fig 7 than in Fig. 5, indicating that empirical cascades exhibit much more structural diversity at any given size compared to those generated by the model.

These simulation results yield several observations. First, 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).

Second, 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 (25M nodes) they are still not as large nor are

their degree distributions as skewed as the actual Twitter follower graph, which comprises roughly 500M users, the most connected of whom have well over 30M 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 is too simplistic. Presumably, that is, content introduced to Twitter exhibits large differences in intrinsic interestingness, breadth of appeal, and therefore likelihood of being shared. Potentially also introducing such variation into our model would also increase the variation of structural virality at any given size. Nevertheless, it is striking that much of the observed variability in structural virality can be obtained without assuming any variability in β whatsoever.

Finally, it is perhaps surprising that our best model fit is for $R_0 \approx 0.7$, a value squarely in a regime traditionally characterized as "subcritical", and one that is not usually associated with viral propagation. Naturally, consistency between an empirically observed pattern and the output of a model does not necessarily imply that the model constitutes a causal explanation of the empirical process in question (Ijiri et al. 1977). However, from the perspective of posing the simplest possible model that accounts for the widest range of empirical observations, it remains striking that "subcritical" diffusion is consistent with even the largest events.

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 virality of information cascades. Although our results are restricted to 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 have measured the fine-grain structure of nearly one billion naturally occurring diffusion events in a specific online setting, namely web content spreading on Twitter. In particular, we have identified hundreds of viral cascades—the largest such collection to date—revealing remarkable structural diversity of diffusion events, ranging from broadcast to viral, and

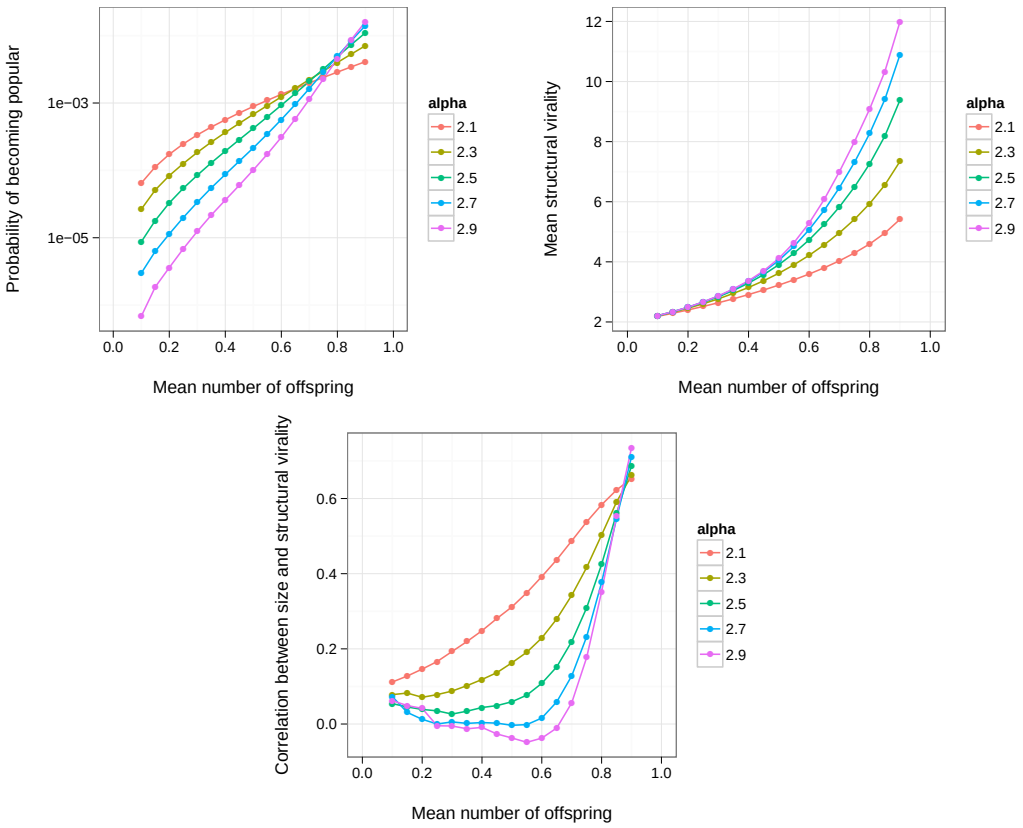

Figure 6: Likelihood of becoming popular (i.e., having at least 100 adopters), mean structural virality, and the correlation between size and structural virality for simulated cascades generated from an SIR model on a random scale-free network, plotted as a function of the model parameters. Each line corresponds to a different exponent α for the power-law network degree distribution, and $R_0 = \beta \bar{d}$ is the expected number of individuals a random node infects in a fully susceptible population.

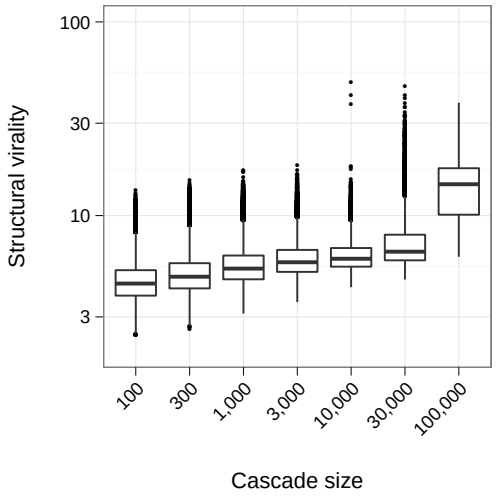

Figure 7: Boxplot of structural virality by size (on a log-log scale) for 1 billion simulated cascades generated from an SIR model on a random scale-free network with $\alpha = 2.3$ and $r = 0.7$.

containing essentially everything in between, where we emphasize that such an exercise would be difficult absent a metric for classifying and ordering many hundreds of thousands of examples automatically. In addition we find relatively low correlation between size and virality, highlighting the difficulty in determining how content spread given only knowledge of its popularity.

Third, we have shown 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 below the root node, or seed (Goel et al. 2012). Clearly events of this latter kind are attributable to subcritical diffusion[16] and hence one might thus be tempted to model online diffusion via two categorically distinct mechanisms, separately accounting for the head and tail 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 this subcritical parameter regime. As discussed earlier, this seems due to the heavy-tailed degree distribution of Twitter, a finding that 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.

Finally, in addition to our three scientific contributions, we note that our work also contributes

[16]For example, Leskovec et al. (2007) found that an SIS model with $\beta = 0.025$, equivalent to $R_0 \approx 0.14$, was able to replicate the size distribution of observed cascades of links over a network of blogs.

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 have only recently become available. That is, only after tracing the propagation of over a billion pieces of content can we collect enough examples 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.

Acknowledgements

The authors are grateful to Ho John Lee for providing access to the Twitter data, and to Robert Gruen for help with the data analysis.

References

Adar, E., L.A. Adamic. 2005. Tracking information epidemics in blogspace. *IEEE/WIC/ACM International Conference on Web Intelligence*. IEEE Computer Society, Compiegne University of Technology, France.

Anderson, Roy M., Robert M. May. 1991. *Infectious Diseases of Humans*. Oxford University Press, Oxford.

Aral, S., L. Muchnik, A. Sundararajan. 2009. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences* **106**(51) 21544–21549.

Bakshy, E., J.M. Hofman, W.A. Mason, D.J. Watts. 2011. Everyone's an influencer: quantifying influence on twitter. *Proceedings of the fourth ACM international conference on Web search and data mining*. Association of Computing Machinery, 65–74.

Bakshy, E., B. Karrer, L.A. Adamic. 2009. Social influence and the diffusion of user-created content. *Proceedings of the tenth ACM conference on Electronic commerce*. Association of Computing Machinery, 325–334.

Barabási, A.L., R. Albert. 1999. Emergence of scaling in random networks. *science* **286**(5439) 509–512.

Bass, Frank M. 1969. A new product growth for model consumer durables. *Management Science* **15**(5) 215–227.

Bass, Frank M. 2004. Comments on a new product growth for model consumer durables the bass model. *Management science* **50**(12 supplement) 1833–1840.

Chierichetti, F., J. Kleinberg, D. Liben-Nowell. 2011. Reconstructing patterns of information diffusion from incomplete observations. NIPS.

Clauset, A., C.R. Shalizi, M.E.J. Newman. 2009. Power-law distributions in empirical data. *SIAM review* **51**(4) 661–703.

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

Dodds, P.S., D.J. Watts. 2004. Universal behavior in a generalized model of contagion. *Physical Review Letters* **92**(21) 218701.

Dow, P Alex, Lada A Adamic, Adrien Friggeri. 2013. The anatomy of large facebook cascades .

Fichman, Robert G. 1992. Information technology diffusion: a review of empirical research. *ICIS*. 195–206.

Goel, S., D.J. Watts, D.G. Goldstein. 2012. The structure of online diffusion networks. *Proceedings of the 13th ACM Conference on Electronic Commerce*. ACM, 623–638.

Golub, B., M.O. Jackson. 2010. Using selection bias to explain the observed structure of internet diffusions. *Proceedings of the National Academy of Sciences* **107**(24) 10833–10836.

Granovetter, M. 1978. Threshold models of collective behavior1. *American Journal of Sociology* **83**(6) 1420–1443.

Ijiri, Yuji, Herbert Alexander Simon, Charles P Bonini, Theodore A van Wormer. 1977. *Skew distributions and the sizes of business firms*. North-Holland Publishing Company New York.

Iyengar, R., C. Van den Bulte, T. W. Valente. 2010. Opinion leadership and social contagion in new product diffusion. *Marketing Science* .

Kempe, David, Jon Kleinberg, Eva Tardos. 2003. Maximizing the spread of influence through a social network. *9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association of Computing Machinery.

Kermack, W.O., A.G. McKendrick. 1927. Contributions to the mathematical theory of epidemics – I. *Proceedings of the Royal Society* **115A** 700–721.

Leskovec, J., L.A. Adamic, B.A. Huberman. 2007. The dynamics of viral marketing. *ACM Transactions on the Web* **1**(1) 5.

Leskovec, J., A. Singh, J. Kleinberg. 2006. Patterns of influence in a recommendation network. *Advances in Knowledge Discovery and Data Mining* 380–389.

Liben-Nowell, D., J. Kleinberg. 2008. Tracing information flow on a global scale using internet chain-letter data. *Proceedings of the National Academy of Sciences* **105**(12) 4633.

Lloyd, Alun L, Robert M May. 2001. How viruses spread among computers and people. *Science* **292**(5520) 1316–1317.

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

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

Mahajan, Vijay, Robert A Peterson. 1985. *Models for innovation diffusion*, vol. 48. Sage.

Mohar, Bojan, Toma Pisanski. 1988. How to compute the wiener index of a graph. *Journal of Mathematical Chemistry* **2** 267–277. URL `http://dx.doi.org/10.1007/BF01167206`. 10.1007/BF01167206.

Newman, M.E.J. 2005. Power laws, Pareto distributions and Zipfs law. *Contemporary physics* **46**(5) 323–351.

Pastor-Satorras, R., A. Vespignani. 2001. Epidemic spreading in scale-free networks. *Physical review letters* **86**(14) 3200–3203.

Rogers, E.M. 1962. *Diffusion of innovations*. Free Press.

Shalizi, C. R., A. C. Thomas. 2011. Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods and Research* **40** 211–239.

Sun, E., I. Rosenn, C. Marlow, T. Lento. 2009. Gesundheit! modeling contagion through facebook news feed. *Proc. of International AAAI Conference on Weblogs and Social Media*.

Toole, Jameson L, Meeyoung Cha, Marta C González. 2012. Modeling the adoption of innovations in the presence of geographic and media influences. *PloS one* **7**(1) e29528.

Valente, Thomas W. 1995. *Network Models of the Diffusion of Innovations (Quantitative Methods in Communication Series)*. Hampton Press (NJ)(January 10, 1995).

Van den Bulte, Christophe, Gary L Lilien. 2001. Medical innovation revisited: Social contagion versus marketing effort1. *American Journal of Sociology* **106**(5) 1409–1435.

Walther, Joseph B, Caleb T Carr, Scott Seung W Choi, David C DeAndrea, Jinsuk Kim, Stephanie Tom Tong, Brandon Van Der Heide. 2010. Interaction of interpersonal, peer, and media influence sources online. *A networked self: Identity, community, and culture on social network sites* **17** 17–38.

Watts, Duncan J. 2002. A simple model of information cascades on random networks. *Proceedings of the National Academy of Science, U.S.A.* **99** 5766–5771.

Wiener, Harry. 1947. Structural determination of paraffin boiling points. *Journal of the American Chemical Society* **69**(1) 17–20. doi:10.1021/ja01193a005. URL `http://pubs.acs.org/doi/abs/10.1021/ja01193a005`.

Wu, Shaomei, Jake M Hofman, Winter A Mason, Duncan J Watts. 2011. Who says what to whom on twitter. *Proceedings of the 20th international conference on World wide web*. ACM, 705–714.

Young, H. Peyton. 2009. Innovation diffusion in heterogeneous populations: Contagion, social influence, and social learning. *American Economic Review* **99**(5) 1899–1924.

Appendix A. Computing structural virality

The average distance measure of structural virality 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,*

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

In particular,

$$\text{dist}_{\text{avg}} = \frac{2n}{n-1}\left[\frac{1}{n}\sum_{S\in\mathcal{S}} |S| - \frac{1}{n^2}\sum_{S\in\mathcal{S}} |S|^2\right]. \tag{3}$$

Proof. Statement (3) in the theorem follows directly from (2) together with Lemma 1, and so we only need to establish statement (2). 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

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

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

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

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

$$\begin{aligned}\sum_{i,j=1}^{n} \mathrm{dist}(v_i, v_j) &= 2n^2 + 2n \sum_{i=1}^{n} \mathrm{depth}(v_i) - 2 \sum_{i,j=1}^{n} \sum_{S \in \mathcal{S}} \delta_S(v_i)\delta_S(v_j) \\ &= 2n^2 + 2n \sum_{i=1}^{n} \mathrm{depth}(v_i) - 2 \sum_{S \in \mathcal{S}} |S|^2.\end{aligned}$$

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

```
1:  function SUBTREE-MOMENTS(T, r)
2:      if T.size() = 1 then                                   ▷ The base case
3:          size ← 1
4:          sum-sizes ← 1
5:          sum-sizes-sqr ← 1
6:      else                                  ▷ Recurse over the children of the root r
7:          for c ∈ r.children() do
8:              size_c, sum-sizes_c, sum-sizes-sqr_c ← SUBTREE-MOMENTS(T, c)

9:          size ← 0
10:         sum-sizes ← 0
11:         sum-sizes-sqr ← 0
12:         for c ∈ r.children() do
13:             size ← size + size_c
14:             sum-sizes ← sum-sizes + sum-sizes_c
15:             sum-sizes-sqr ← sum-sizes-sqr + sum-sizes-sqr_c

16:         size ← size + 1
17:         sum-sizes ← sum-sizes + size
18:         sum-sizes-sqr ← sum-sizes-sqr + size^2

19:     return size, sum-sizes, sum-sizes-sqr

20: function AVERAGE-DISTANCE(T, r)
21:     size, sum-sizes, sum-sizes-sqr ← SUBTREE-MOMENTS(T, r)
22:     dist_avg ← [2 · size/(size − 1)] [sum-sizes/size − sum-sizes-sqr/size^2]

23:     return dist_avg
```

Table 1: Rank correlation between alternative measures of structural virality

	Average Distance	Relative Broadcast	Distinct Parent	Average Depth
Average Distance	1	-0.92	0.94	0.94
Relative Broadcast	-0.92	1	-0.99	-0.89
Distinct Parent	0.94	-0.99	1	0.91
Average Depth	0.94	-0.89	0.91	1

Appendix B. Alternative measures of structural virality

Although we have demonstrated that our particular definition of structural virality is quite effective, 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 multi-generational "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 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 0.92 with all three alternative measures, and also, all pairwise rank correlations are at least 0.89. Moreover, our empirical results are qualitatively similar regardless of which of these alternative measures of structural virality we apply. For example, Fig. 8 shows the relationship between size and the probability two nodes have different parents.

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. Off-channel diffusion and homophily

While 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

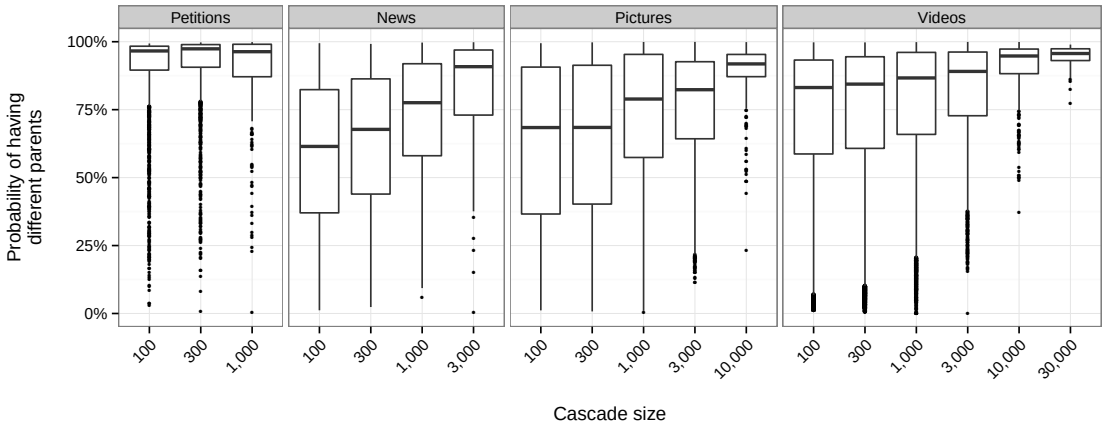

Figure 8: Boxplot of an alternative measure of structural virality—probability of two nodes having different parents—by size (on a log scale), separated by domain. Lines inside the boxes indicate the median, while the boxes themselves show interquartile ranges.

also spread via other channels, such as Facebook or email. As a result of this "off-channel" diffusion, two individuals on Twitter who appear 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 only considering 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 relatively to hashtags in general, they are still plentiful, amounting to 94,000 cascades with at least 100 adopters. Figs. 9 and 10 show that the diffusion of these long hashtags yield qualitatively similar results to our primary analysis, suggesting that off-channel diffusion is not driving our findings.

To account for the potential confounding of influence and homophily, we repeat our analysis after limiting to adoptions that explicitly reference the original root node of their cascade (a by-product of using Twitter's official retweet function), mitigating the likelihood that two distinct cascades are

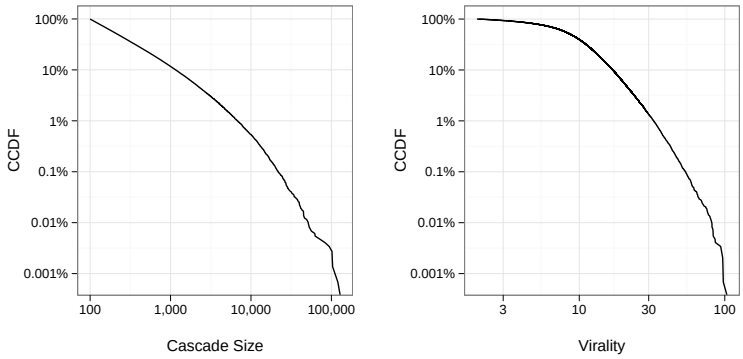

Figure 9: Size and structural virality distributions on a log-log scale for popular hashtag cascades, containing at least 100 adopters.

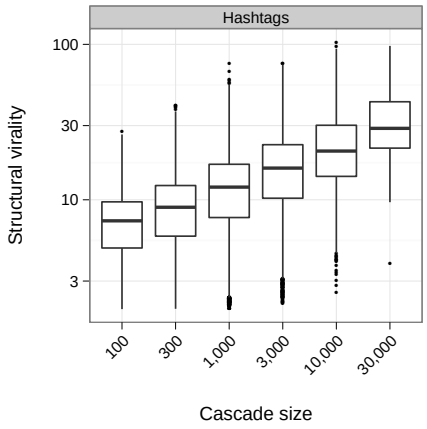

Figure 10: Boxplot of structural virality by size on a log-log scale for hashtag cascades. Lines inside the boxes indicate median structural virality, while the boxes themselves show interquartile ranges.

mistakenly merged.[17] As with the off-channel results above, Figs. 11 and 12 show that the revised analysis produces results consistent with our primary analysis, thereby diminishing concerns that our results are a consequence of confounding.[18]

[17]Because of the large number of tweets, we limit to a 1% sample, constituting 203,137 cascades having at least 100 adopters.

[18]We also note that if present, a homophily bias would lead us to overestimate the size of diffusion trees, in contrast with off-channel diffusion which would have the opposite effect.

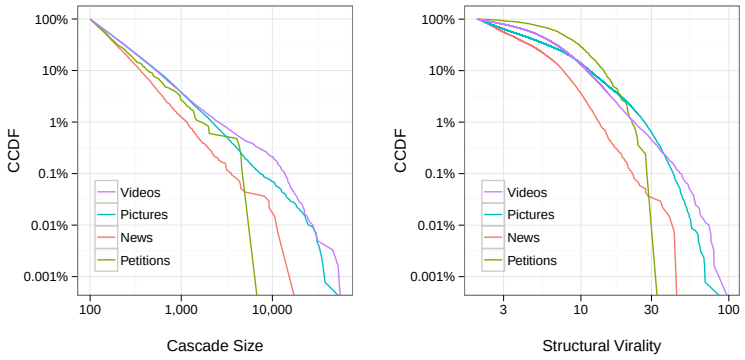

Figure 11: Size and structural virality distributions on a log-log scale for popular official retweet cascades, containing at least 100 adopters, separated by domain.

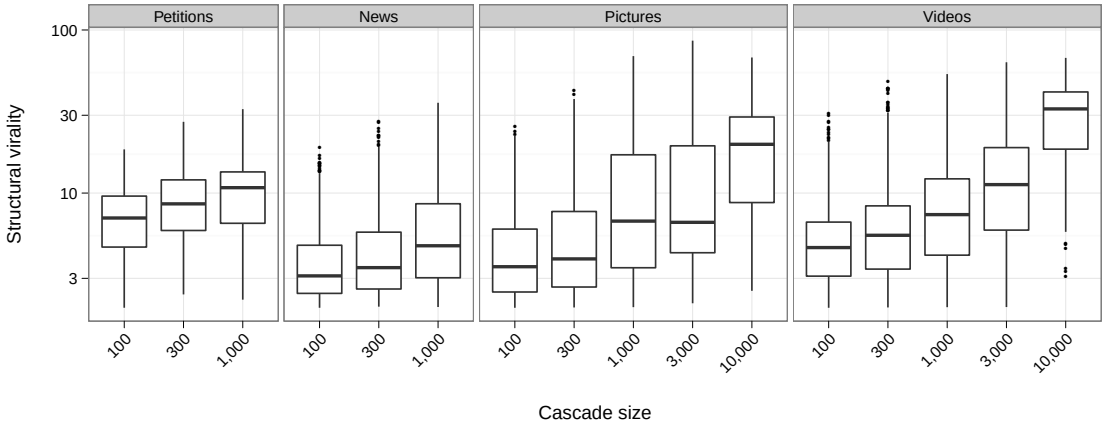

Figure 12: Boxplot of structural virality by size on a log-log scale for official retweet cascades, separated by domain. Lines inside the boxes indicate median structural virality, while the boxes themselves show interquartile ranges.