The Structure of Online Diffusion Networks

SHARAD GOEL, Yahoo! Research
DUNCAN J. WATTS, Yahoo! Research
DANIEL G. GOLDSTEIN, Yahoo! Research

Models of networked diffusion that are motivated by analogy with the spread of infectious disease have been applied to a wide range of social and economic adoption processes, including those related to new products, ideas, norms and behaviors. However, it is unknown how accurately these models account for the empirical structure of diffusion over networks. Here we describe the diffusion patterns arising from seven online domains, ranging from communications platforms to networked games to microblogging services, each involving distinct types of content and modes of sharing. We find strikingly similar patterns across all domains. In particular, the vast majority of cascades are small, and are described by a handful of simple tree structures that terminate within one degree of an initial adopting "seed." In addition we find that structures other than these account for only a tiny fraction of total adoptions; that is, adoptions resulting from chains of referrals are extremely rare. Finally, even for the largest cascades that we observe, we find that the bulk of adoptions often takes place within one degree of a few dominant individuals. Together, these observations suggest new directions for modeling of online adoption processes.

Categories and Subject Descriptors: J.4 [**Social and Behavioral Sciences**]: Sociology

General Terms: Economics, Measurement

Additional Key Words and Phrases: Computational social science, contagion, diffusion, social networks

1. INTRODUCTION

A longstanding hypothesis in diffusion research is that adoption of products and ideas spreads through interpersonal networks of influence analogous to the manner in which an infectious disease spreads through a susceptible population [Anderson and May 1991]. Accordingly, theoretical models of diffusion have generally imitated disease models in the sense that popular products are assumed to diffuse multiple steps from their origin in the manner of epidemics, "infecting" large numbers of people in the process [Watts 2002; Leskovec et al. 2006]. Although this assumption is entirely plausible, empirical diffusion research has historically relied on aggregate data, such as cumulative adoption curves [Coleman et al. 1957; Bass 1969; Young 2009; Iyengar et al. 2010], which reveal only the total number of adopters at any given time. While these curves are consistent with the hypothesis of "viral", disease-like diffusion, they are also consistent with other mechanisms, such as marketing or mass media [Van den Bulte and Lilien 2001]. As a consequence, the extent to which adoption processes are driven by these different mechanisms, and therefore the extent to which prevailing models accurately describe online diffusion, is unknown.

In recent years, the increased availability of online social interaction data has offered new opportunities to map out the network structure of diffusion processes [Adar

Author's addresses: Microeconomics and Social Systems, Yahoo! Research, 111 West 40^{th} Street, New York, NY 10018. Correspondence may be sent to S.G. at goel@yahoo-inc.com.

EC'12, June 4–8, 2012, Valencia, Spain.

and Adamic 2005; Leskovec et al. 2006, 2007; Bakshy et al. 2009; Sun et al. 2009; Bakshy et al. 2011]. Here we leverage online diffusion data to address the question of how much total adoption derives from viral spreading versus some other process, as well as the implications of this empirical result for theoretical models of diffusion, in particular online diffusion.

In order to identify generic features of online diffusion structure, we study seven diverse examples comprising millions of individual adopters. As opposed to biological contagion, our domain of interest comprises the diffusion of adoptions, where "adoption" implies a deliberate action on the part of the adopting individual. In particular, we do not consider mere exposure to an idea or product to constitute adoption. Contagious processes such as email viruses, which benefit from accidental or unintentional transmission are therefore excluded from consideration.

Although restricted in this manner, the range of applications that we consider is broad. The seven studies described below draw on different sources of data, were recorded using different technical mechanisms over different timescales, and varied widely in terms of the costliness of an adoption. This variety is important to our conclusions, as while each individual study no doubt suffers from systematic biases arising from the particular choice of data and methods, collectively they are unlikely to all exhibit the same systematic biases. To the extent that we observe consistent patterns across all examples, we expect that our findings should be broadly applicable to other examples of online—and possibly offline—diffusion as well.

The remainder of this paper proceeds as follows. After reviewing the diffusion literature in Section 2, in Section 3 we describe in detail the seven domains we investigate. We present our main results in Section 4, showing that not only are most *cascades* small and shallow, but also that most *adoptions* lie in such cascades. In particular, it is rare for adoptions to result from chains of referrals. Finally, in Section 5 we discuss the implications of these results for diffusion models, as well as the apparent discord between our results and the prevalence of popular products, such as Facebook and Gmail, whose success is often attributed to viral propagation.

2. RELATED WORK

The adoption of new products or behaviors has been described rhetorically in the language of contagion for centuries [Mackay 1841; Le Bon 1896]. By the late 1960's, the conceptual link between adoption and diffusion had been greatly reinforced by their enshrinement in simple mathematical models, which were derived by analogy either from mass-action laws of chemical reactions [Coleman et al. 1957], or from classical models of mathematical epidemiology [Bass 1969]. These various models, however, all embodied the core substantive assumption that new adopters are influenced by the proportion of the population that has adopted previously. The main empirical prediction of these models was that viewed over time, the cumulative number of adopters in a population ought to be described by an "S shaped" curve—a prediction that was at least broadly consistent with a large number of empirical studies [Rogers 1962; Bass 1969].

Subsequently a large literature has developed in marketing and related fields that has explored the mathematical properties of disease-like models. In addition, related but distinct families of models have emerged that begin from different assumptions about the psychology of the adoption process. Granovetter [1978] and others [Lopez-Pintado and Watts 2008], for example, have argued that many adoption decisions—especially costly ones—are more appropriately modeled as a nonlinear "threshold," whereupon adoption takes place only after some critical number or fraction of some sample population had adopted. Dodds and Watts [2005] have proposed a model of "generalized contagion" that contained both disease-like and threshold models as spe-

cial cases, while Young [2009] has proposed a model of observational learning that also exhibits threshold-like behavior. Although they differ in their psychological motivation and formal properties, a common feature to all these models is that a small "seed" set of initial adopters—possibly just one—can, under the right circumstances, trigger a much larger "cascade" of subsequent adoptions that builds upon itself over multiple, possibly many, generations.

The dynamics and precursors of large cascades, or "social epidemics" as they are also known [Gladwell 2000], have also been studied extensively in networked models of adoption, where the classical uniform mixing assumption—namely that the probability of adoption depends only on the global ratios of adopters and non-adopters [Anderson and May 1991]—is replaced with the assumption that individuals are influenced only by some relatively small number of network neighbors [Moore and Newman 2000; Watts 2002]. An important contribution of these network diffusion models is that unlike in the case of uniform mixing, for which all seeds are effectively interchangeable, both the local and global structural properties of networks can greatly influence the size and likelihood of a cascade that is triggered by any given seed.

Motivated by this observation, a growing number of papers have addressed the natural question of how seeds can be selected so as to maximize the total amount of influence that some exogenous agent can hope to exert over a network with some given structure. Assuming a simple variant of a network influence model, for example, Domingos and Richardson [2001], estimated the "viral lift" that could be attained by incentivizing a single consumer to adopt. Subsequently, Kempe et al. [2003] found approximately optimal algorithms for selecting a maximally influential set of seeds for a variety of simple models. And recently Kitsak et al. [2010] showed that for simple disease-like models of adoption, diffusion was maximized when initiated by individuals in the dense core of the network.

Whereas much of the earlier literature focuses on the conditions required for a large cascade to occur at all, the influence maximization literature focuses on the operational question of how to efficiently trade off between the cost of seeding and the size of the cascade generated. Nevertheless, both literatures start from the assumption that a relatively small number of seeds can trigger a relatively large number of adoptions via some, usually multistep, diffusion process. In their numerical experiments, for example, Kempe et al. [2003] find that depending on the assumed infection probability, optimal targeting can generate cascades from several times to several hundred times the size of the seed set. Domingos et al. [2005] report an even more extreme result, that the viral lift from a single targeted consumer can be as high as 20,000 others. Finally, many models of adoption [Bass 1969; Granovetter 1978; Watts 2002] find that under the right circumstances, cascades triggered by a single seed can encompass the entire population, regardless of size.

Whether it is stated explicitly or not, in other words, the assumption that *a relatively small number of seeds can trigger a relatively large number of adoptions via some, usually multistep, diffusion process* is central to many of the interesting questions posed by the modeling literature. But this assumption provokes two related empirical questions: how frequently do such large cascades occur in real diffusion processes; and how much total adoption do they account for? The central contribution of this paper is to directly address these two questions in the context of online diffusion.

Answering these questions has been difficult for offline adoption processes, for which the data have historically suffered from two limitations. First, most empirical diffusion studies have relied on aggregated data, reflecting the total number of adoptions in a population. As has been pointed out elsewhere [Van den Bulte and Lilien 2001], although these adoption curves are consistent with diffusion processes, they are also consistent with alternative explanations such as marketing or advertising efforts, or

simply heterogeneity in adoption propensity. To address this problem, individual-level diffusion data are required, meaning that one can observe precisely who influences whom to adopt which product at a given time. In an offline context, however, data of this kind are extremely difficult to collect, especially at the scale required to study large cascades spreading over potentially many generations of adopters. A second problem with empirical diffusion studies has been that they are highly subject to selection bias, as examples of successful diffusion are much easier to observe than examples of unsuccessful ones [Liben-Nowell and Kleinberg 2008].[1] More generally, without a representative sample of diffusion *attempts*, it is difficult to quantify the overall rate at which diffusion takes place or how much total adoption it accounts for.

A major advantage of studying diffusion in an online context, therefore, is that it is increasingly possible to overcome these limitations, allowing researchers first, to obtain very large samples of potential online diffusion—for example all news stories published on Twitter, all videos posted to YouTube, or all third-party applications launched on Facebook—regardless of whether they actually diffuse or not; and second, to map out the precise network structure of the corresponding diffusion processes as they propagate from individual to individual.

In recent years, a number of studies [Adar and Adamic 2005; Bakshy et al. 2009; Sun et al. 2009; Bakshy et al. 2011] have made use of online data to study various diffusion processes. As we discuss later, the results of these studies are consistent with ours; however, they have not directly addressed the central question of how much total adoption derives from viral spreading versus some other process. Of particular relevance is Leskovec et al. [2006, 2007], who analyzed product recommendations in a network of users of an e-commerce website with the primary objective of enumerating and counting the types of diffusion cascades that arose. As we do, they find that most cascades are small; however, they do not consider the subsequent—and from our perspective key—question of how much adoption is accounted for by the minority of large cascades.

In addition to this different objective, we note that Leskovec et al. [2006, 2007] investigate one domain with a high barrier to adoption: users must receive a recommendation and purchase a product to be counted as adopters. In contrast, we compare seven domains in which the type and the cost of adoption ranges greatly, from "retweeting" on Twitter to sending an email to purchasing. Furthermore, our analysis covers domains in which there is ambiguity as to which is the parent node in the cascade (as in the recommendations domain of Leskovec et al. [2006, 2007]) but also domains in which parents are uniquely identified through tracking URLs. The larger point is that because each prior study of online diffusion focused on a single web platform, such as Facebook [Sun et al. 2009], Twitter [Bakshy et al. 2011], Second Life [Bakshy et al. 2009], or blog networks [Adar and Adamic 2005] and also invoked different metrics, it has previously been difficult to draw conclusions about online diffusion processes in general.

3. DATA AND METHODS

For each domain, we define a diffusion event or "cascade" [Watts 2002; Kempe et al. 2003] as comprising a "seed" individual, who takes the relevant adoption action independently of any other individual in our dataset, followed by other non-seed individuals who are influenced either directly or indirectly by the seed to take the same action. From this definition, it follows that every individual in a diffusion event can be

[1]Even modeling efforts have tended to focus on the mechanics of successful diffusion rather than on its likelihood. For example, the Bass model deterministically predicts universal diffusion, and is hence unable to account for attempts at diffusion that do not succeed.

connected by some unbroken path of adoptions back to the original seed, hence each diffusion event can be represented as a single connected graph. Moreover, in instances where an adopter may have been influenced by more than one previous adopter, we ascribe influence exclusively to the earliest such "parent" node; thus in contrast with related work [Leskovec et al. 2007], where multiple parents are allowed, we map all diffusion events to tree structures that originate with a single seed and terminate at one or more "leaf" nodes.[2] Finally, having reconstructed all distinct diffusion events, we calculate two key quantities: first, the frequency of distinct diffusion structures (trees) that we observe, and second, the proportion of total adoptions that are accounted for by each of these structures. Given the scale of the data, calculations are carried out with the MapReduce parallel computation framework [Dean and Ghemawat 2008].

Before describing the data in detail, we note that although the definitions just outlined are consistent across all seven domains, the manner by which we infer influence differs across them. In three of the applications, we directly observe interpersonal diffusion (influence), whereas in the remaining four we infer influence from the underlying network of interpersonal connections and the temporal sequence of adoptions. Although in the latter case it is possible that apparent influence can be accounted for simply by homophily [Aral et al. 2009; Shalizi and Thomas 2011], our inferred diffusion trees nevertheless provide an effective upper bound on the actual diffusion taking place via the respective underlying networks. Moreover, while it is possible that some amount of diffusion goes undetected (e.g., if it occurs over a channel, such as literal word-of-mouth, that we cannot track), given the relative ease with which one can share content via Twitter, Facebook, email and other modes of electronic communication, we suspect this unobserved adoption accounts for only a small fraction of total adoptions of the online products investigated here.

3.1. Observed-Diffusion Domains

(1) *Yahoo! Kindness*. Over a one month period in 2010, Yahoo!'s philanthropic arm launched a website (kindness.yahoo.com) that asked users to create status updates describing kind acts they had performed, after which these updates were propagated via Yahoo!, Facebook, Twitter, and other means in order to attract new users to visit the site and post updates of their own. Because all users were logged in, and because each user received a unique coded URL when arriving at the site (e.g., kindness.yahoo.com/1QvTu) that was used to bring others to the site, it was possible to trace the chain of adopters through which each new user arrived. Because this tracking method works regardless of how people chose to share links (e.g., by email, blog, instant message, status update, forum, etc.), we could reconstruct the diffusion of the new site across the Internet. During the course of the experiment, approximately 59,000 users "adopted" the campaign, meaning that they visited the site and also posted at least one status update.

(2) *Zync* [Liu et al. 2007; Shamma and Liu 2009] is a plug-in for Yahoo! Messenger, an instant messaging (IM) application, that allows pairs of users to watch videos synchronously while sending instant messages to one another. Individuals initiate a session by sending a video URL to another user through the Yahoo! instant messaging client. The invited user must accept the invitation before the video commences; thus in contrast with our other examples, a single use of Zync requires two users. Since 2009, Zync has been activated by approximately 1.3 million users; however, to avoid counting spurious dyads, we define adoption in this instance

[2]Given the structural simplicity of the vast majority of cascades we find—detailed in Section 4—our results are qualitatively the same regardless of which particular cascade representation is chosen.

as having initiated a session, not merely having accepted an invitation, yielding 374,000 adopters.

(3) *The Secretary Game* is an online variant of the "secretary problem" [Ferguson 1989], a sequential search game in which players attempt to find an optimal stopping rule. Players are encouraged to share the game's URL with at least three other people with an explanation that the game designers are seeking the world's best players. As with Yahoo! Kindness, user-specific URLs tracked player-to-player diffusion. Between 2008 and 2011, the game was played over 37,500 times by nearly 2,900 adopters.

3.2. Inferred-Diffusion Domains

In the following cases, we observe time-stamped adoptions occurring over a known network. Here, we did not directly observe interpersonal transmission of information but rather inferred that the "parent" of an adopting node—if one existed—was its first network contact to adopt before it did.

(1) *Twitter News Stories.* We tracked the diffusion of 80,000 news stories posted on the microblogging service Twitter during November 2011, where the original article was distributed by one of five popular news sites: The New York Times, CNN, MSNBC, Yahoo! News, and The Huffington Post. Individuals were said to have "adopted" an article if they posted (i.e., "tweeted") a link to the story. In total, we observed 288,000 adoption events. To mitigate left-censoring, we only counted URLs that had not appeared for two weeks prior to the first adoption we observed (i.e., any previous diffusion must have occurred at least two weeks prior and then completely disappeared only to restart). Because the timescale of diffusion on Twitter lasts only a few days in the vast majority of cases, this approach avoids almost all possible left-censoring. Correspondingly, we addressed the possibility of right-censoring by considering only diffusion events that had been initiated at least two weeks prior to the end of the observation period; all cascades where thus observed for at least two weeks.

(2) *Twitter Videos.* Analogous to the news articles, we tracked 540,000 YouTube videos posted on Twitter during November 2011, where users were again said to have "adopted" a particular video if they tweeted a link to it. 1.3 million adoption events were observed.

(3) *Friend Sense* was a third-party Facebook application that queried respondents about their political views as well as their beliefs about their friends' political views [Goel et al. 2010]. Because the analysis required knowing the friends' actual views, various sharing features were built into the application to encourage viral growth. Over the course of a four-month period in 2008, close to 2,500 individuals used the application, providing over 100,000 answers to more than 80 questions.

(4) *Yahoo! Voice* is a paid service launched in 2004 that allows users to make voice-over-IP calls to phones through Yahoo! Messenger. Between 2004 and 2009, 1.8 million users purchased voice credits, all of whom we define to be adopters. Diffusion in this case is considered to occur over the Yahoo! Messenger IM network, which comprises over 200 million users with a median of 6 network neighbors each, where two individuals are connected if they list each other as a "buddy" (i.e., social contact).

3.3. Tree Canonicalization

In computing the frequency of cascade structures across the diffusion datasets, care needs to be taken to aggregate all variations of the same fundamental structure. For

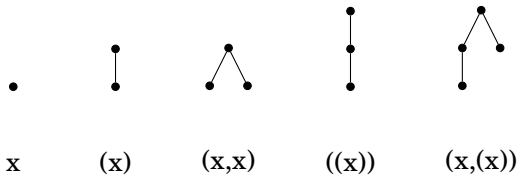

Fig. 1. Canonical names for some example trees.

example, the two trees below, while superficially different, should clearly be taken to represent the same cascade structure for the purposes of our analysis.

In particular, we would like to aggregate all *isomorphic* trees (i.e., trees that are identical under a relabeling of the vertices.)

Definition 3.1 (*Tree Isomorphism*). Consider two rooted trees $T_1 = (V_1, E_1, r_1)$ and $T_2 = (V_2, E_2, r_2)$, with vertices V_i, edges $E_i \subseteq V_i \times V_i$, and roots $r_i \in V_i$. Then T_1 and T_2 are *isomorphic* if there exists a bijection $\phi : V_1 \to V_2$ such that $\phi(r_1) = r_2$ and $(v_1, v_2) \in E_1 \iff (\phi(v_1), \phi(v_2)) \in E_2$.

Determining whether two arbitrary graphs are isomorphic is, in general, a difficult computational problem. No polynomial-time algorithm has been found, and somewhat surprisingly, it is not even know whether the problem lies in either P or NP-complete. In the case of trees, however, there are standard and efficient ways for determining isomorphism, one of which we detail below.

Definition 3.2. The *canonical name* $c(T)$ of a rooted tree T is a string defined inductively on the height of the tree by the following two rules:

(1) (Basis) The canonical name for the one-node tree is **x**.
(2) (Induction) If T has more than one node, let $T_1, \ldots, T_k$ denote the subtrees of the root indexed such that $c(T_1) \leq c(T_2) \leq \cdots \leq c(T_k)$ under the lexicographic order. Then the canonical name for T is

$$\mathbf{(}c(T_1)\mathbf{,}\ldots\mathbf{,}c(T_k)\mathbf{)}.$$

Figure 1 shows the canonical names for a few example tree structures. It is clear that two trees have the same canonical name if and only if they are isomorphic. Moreover, as shown in Aho et al. [1974], $c(T)$ can be computed in time linear in the number of nodes. These canonical tree names thus allow us to efficiently aggregate all isomorphic variations of each cascade structure.

4. RESULTS

Before presenting our results, we note that in addition to variations in the method of data collection, our examples also varied widely with respect to a number of dimensions thought to be important to adoption decisions, such as the costliness of the adoption, the nature of the network over which the adoptions are diffusing, and the timescale on which the diffusion process proceeded. For example, whereas Yahoo! Voice represents an adoption decision that costs money, adoptions in the gaming domains are costly only in time, and Twitter retweets are next to costless. Whereas the IM network represents a network of reciprocated, private communication ties, the majority of Facebook edges do not involve active communication, and edges on Twitter are largely

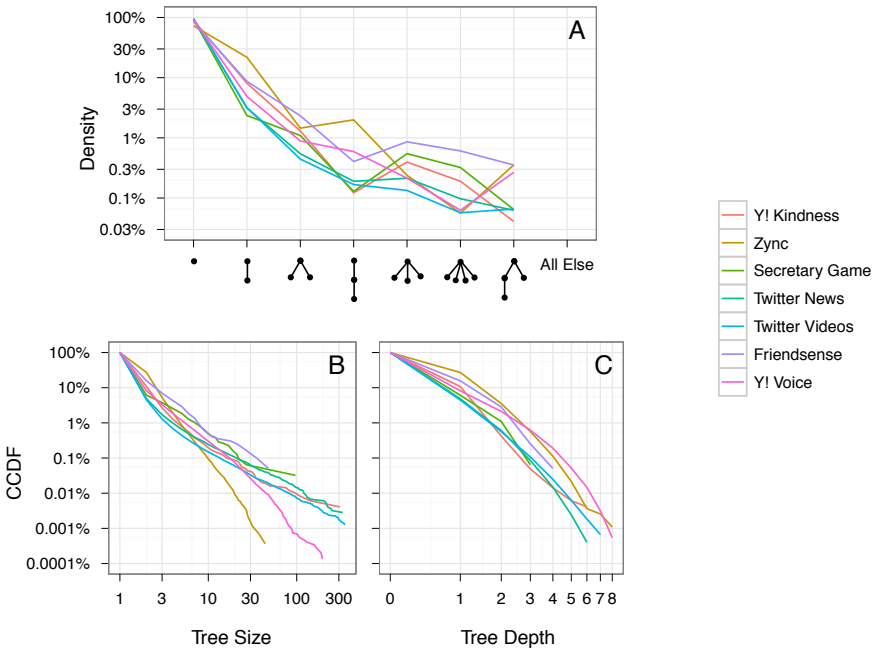

Fig. 2. The distribution of diffusion cascade structures

unreciprocated and public. Whereas Zync, Friend Sense, and Yahoo! Voice clearly exhibit positive externalities in the sense that the utility of the products in question increases with the number of adopting neighbors, such network effects are less likely in the remaining domains. And whereas the diffusion cascades on Twitter generally terminated within a day or two, the Secretary Game and Friend Sense spread actively for several weeks, while cascades on Yahoo! Voice extended over several years.

Given the heterogeneity in data collection, timescales (ranging from days to years), and the nature of adoptions described above, the distribution of diffusion structures across all seven cases is striking in its similarity. Fig. 2A shows the frequency of cascades accounted for by the most commonly occurring tree structures across the seven domains we study. The vast majority of instances—ranging from 73% to 95% across domains—show no diffusion at all (i.e., the tree consists only of the seed), while the next most frequent outcome is in all cases a single additional adopter. In fact, the same seven simple tree structures account for upwards of 97% of cascades in each domain. Figs. 2B and 2C complement this result, showing that the distributions of tree size and depth, respectively, are likewise extremely skewed. In all domains, less than 1% of cascades consist of more than seven nodes, and less than 4% extend further than one degree from the seed node.

Although the similarity across domains is striking, our finding that most cascades are small and shallow is not, on its own, surprising. A number of recent empirical studies of online diffusion [Adar and Adamic 2005; Leskovec et al. 2007; Bakshy et al. 2009; Sun et al. 2009; Bakshy et al. 2011] have also observed that the size distribution of diffusion events is right-skewed and heavy-tailed, which necessarily implies that most events are small; indeed, Leskovec et. al [Leskovec et al. 2007] even identify many of the same motifs. The usual intuition regarding heavy-tailed distributions, however, is that large events, although rare, are sufficiently large to dominate certain key proper-

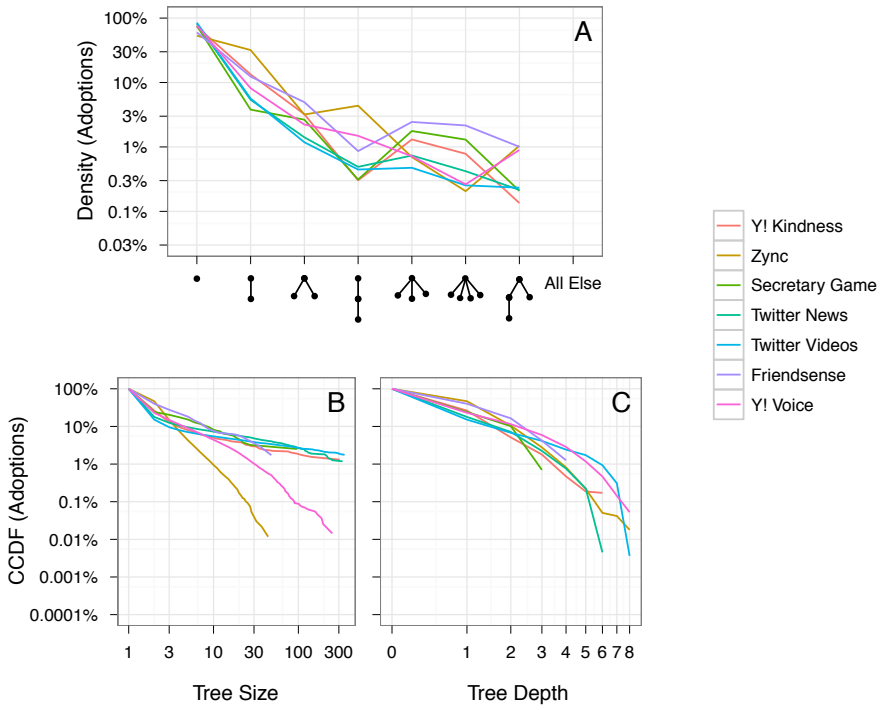

Fig. 3. The distribution of adoptions across cascade structures (i.e., the fraction of adopters residing in each tree type).

Table I. Summary diffusion statistics.

Domain	% of One-Node Cascades	Mean Cascade Size	Mean Cascade Depth	% of Adoptions Within One Degree of a Seed
Yahoo! Kindness	89%	1.2	0.1	99%
Yahoo! Zync	73%	1.4	0.3	96%
The Secretary Game	94%	1.2	0.1	97%
Twitter News Stories	95%	1.2	0.1	98%
Twitter Videos	96%	1.1	0.1	96%
Friend Sense	84%	1.4	0.2	94%
Yahoo! Voice	92%	1.2	0.1	94%

ties of the corresponding system. To illustrate, it is likely that in the course of human history avian influenza has jumped from birds to humans hundreds or even thousands of times, and that the vast majority of such events have led to only one or at most a few infections, much as we see here. Nevertheless, the vast majority of *infections* belong to a handful of very large cascades, which are the epidemics of historical record. Arguably, the majority of all humans who have ever been infected with avian influenza were infected during a single event, namely the 1918 pandemic, during which more than 500 million individuals are thought to have been infected [Taubenberger and Morens 2006]. In our domain, therefore, even if it is the case that 99% of cascades are small, if it is also the case that the remaining 1% are extremely large, epidemic-like trees, the large cascades could still account for the bulk of all adoption activity.

What is surprising, therefore, is that we find no evidence of such epidemics in our data. On the contrary, Fig. 3A shows that the seven tree structures from Fig. 2A also

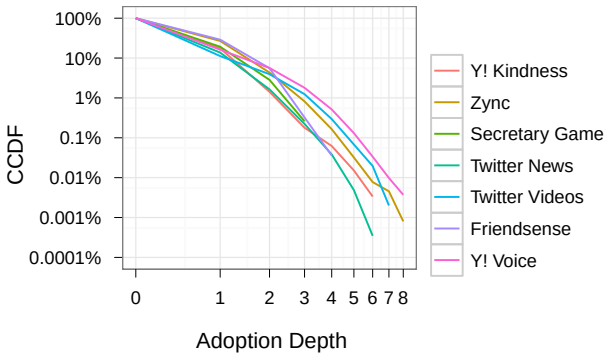

Fig. 4. The distribution of adoptions by depth, which indicates that the vast majority of adoptions occur within 1 generation of a seed.

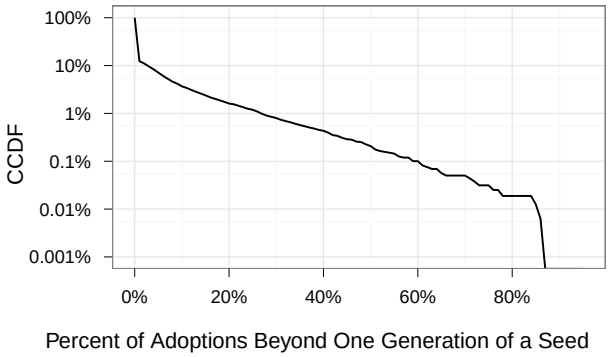

Fig. 5. For the approximately 16,000 videos and news stories posted on Twitter independently by at least ten users, the distribution of the proportion of adoptions occurring beyond one generation of a seed. In particular, less than 1% of these "products" have the majority of their adoptions occurring beyond one generation of a seed.

account for around 90% of all *adoptions*, and thus only a minority of adoptions are found within large cascades. Fig. 3B, in fact, shows that in each of the seven domains less than 10% of adoptions occur in cascades consisting of more than 10 nodes, and similarly Fig. 3C shows that less than 10% occur in trees that extend more than two generations from the seed. Finally, Fig. 4 demonstrates that very few adoptions (1%–6% across domains) take place more than one degree from a seed node, regardless of the size or depth of the tree in which they occur. In other words, in contrast with the intuition of viral spread leading to rare but large, multi-step epidemics, we find that the vast majority of adoptions occur either without peer-to-peer influence or within one step of such an independent adopter. Large cascades, that is, are not only rare, but are also insufficiently large to appreciably alter average cascade size, which varies within a remarkably narrow range of 1.1–1.4 across domains (Table 1).

Although we have considered a diverse set of examples of online diffusion, where in all cases the platform designers (or message initiators in the case of Twitter) had the intention of promoting the diffusion of their "products", it is possible that we have simply not studied enough distinct instances to witness a case of truly viral spread, in

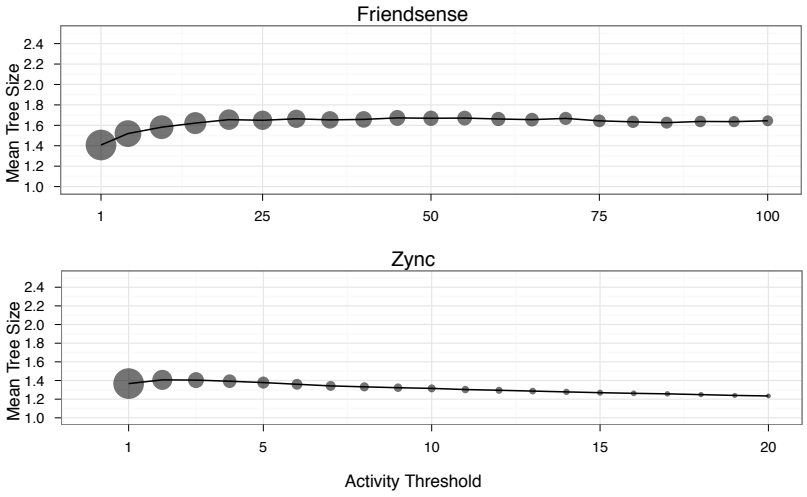

Fig. 6. The effect of varying the threshold for adoption on mean tree size. The area of each circle indicates the proportion of individuals meeting the corresponding activity threshold, with the leftmost points ($k = 1$) comprising all adopters, and the rightmost consisting of less than 10% of adopters in each domain.

which case our results may understate the importance of rare but large diffusion cascades. Ultimately this objection is difficult to refute, as no matter how many negative cases are presented, it is always possible that the sample is neither sufficiently large nor representative to exhibit the feature of interest. Nonetheless, we can address this objection by dramatically increasing the number of cases considered.

Revisiting the Twitter data, from the total sample of over 620,000 videos and news articles, 16,000 were introduced independently by at least 10 individuals. Treating each one of these 16,000 distinct pieces of content as a "product", we now investigate the diffusion structure for this much larger product sample, where each experiences many opportunities to spread virally. As we show in Fig 5, less than 1% of these videos and news stories have the majority of their adoptions occurring beyond one degree from the seed, and none have at least 90% of adoptions beyond one degree. In other words, it is relatively rare for news and videos on Twitter to satisfy even a modest definition of "viral", where once again almost all adoptions take place within one degree of the seed nodes. To the critique that we may have previously considered only unattractive (and thus non-viral) products, we observe that this analysis focuses on the ostensibly most interesting videos and news stories—those submitted by ten or more people—and nonetheless finds cases of viral diffusion to be rare.

Finally, one might object that our definition of adoption biases our conclusions by including individuals as adopters even when they have only minimally interacted with the given product. In particular, among "true converts", who are arguably more likely to discuss, praise and spread products, a qualitatively different picture may emerge in which most adoptions are encompassed within a few large cascades. To investigate this possibility, we exploit a feature shared by two of our examples–Friend Sense and Zync—that allows us to vary the definition of adoption systematically, by counting as adoptions only those instances in which the application in question was played or used at least k times, where $k = 1$ corresponds to the previous analysis, and large k would count only the most enthusiastic adopters. As Fig. 6 shows, the average tree size is

largely invariant with respect to adoption threshold, suggesting that our conclusions are also robust with respect to particular definitions of adoption.

5. DISCUSSION

Our observation that multi-step diffusion is not only rare in these online domains, but that the vast majority of adoptions occurs within one degree of a seed node, has several implications for theories of diffusion as they apply to online content and possibly for adoption processes more generally. First, although the motifs cataloged in Figs. 2 and 3 are quite likely consistent with those predicted by standard epidemiological models for sufficiently low rates of infectiousness, we would argue that the academic literature on diffusion has paid little attention to this "low infectivity" parameter range, focusing instead on the so-called supercritical parameter regime, which yields large, epidemic-like events that propagate for many generations and "infect" large populations [Bass 1969; Moore and Newman 2000; Watts 2002; Kempe et al. 2003; Liben-Nowell and Kleinberg 2008]. As we have noted, the implicit justification for this focus is that even if epidemic-like events are rare, they are occasionally so large that understanding and predicting them is of both theoretical and practical importance. Our results challenge this conventional wisdom. We find not only that multi-step diffusion occurs rarely on these platforms, but that even when it does, it accounts for only a small percentage of total adoptions. To the extent that such theoretical models of adoption are intended to explain typical diffusion events, we thus advocate more emphasis on subcritical processes.

A second, related, point is that even if one could explicitly encode low transmission rates into diffusion models, we argue that doing so would ignore a key scientific goal, namely identifying the underlying causes of these observed rates. Elaborating on this point, consider a hypothetical model that incorporates just two effects: (1) the propensity for individuals to encounter products outside their immediate social network (e.g., via mass media); and (2) the degree of similarity in product taste between social contacts. The dynamics of such a model would therefore be governed by two countervailing forces. On the one hand, as individuals are exposed to more products through advertising and mass media, adoptions due to social referrals—and hence viral cascades—become less prevalent. On the other hand, as the preferences of network neighbors become more similar, social referrals exhibit more value, increasing the frequency of viral adoptions. Although a formal analysis of a model of this type is left for future work, and although there are no doubt many other factors that one might think to include in such a model, our general claim is that by explicitly modeling the psychological antecedents of adoption, even a relatively simple model could lend insight to our empirical observations beyond simply asserting that they correspond to low-infectivity examples.

Finally, our observation that the vast majority of adoptions we study do not result from multi-step diffusion naturally raises the question of what, then, does account for truly large adoption events that do occur, such as online videos that generate many millions of views in a short period of time, or products like Facebook, Gmail, or Hotmail, the sudden popularity of which is often attributed to viral diffusion over networks of individuals? One possibility, consistent with our results, is that events such as these are not strictly viral at all in the sense implied by infectious-disease models, but rather obtain the bulk of their attention either from traditional advertising or from other coverage by the mass media. Media efforts, that is, might generate a large wave of adopters without requiring any viral, peer-to-peer growth of the sort implied by epidemic models. Indeed, it is precisely a version of this "broadcast diffusion" that we observe for the largest cascades in our data. As shown in Fig. 7, the four largest cascades (all from Twitter) are relatively shallow, spreading only a few steps from the

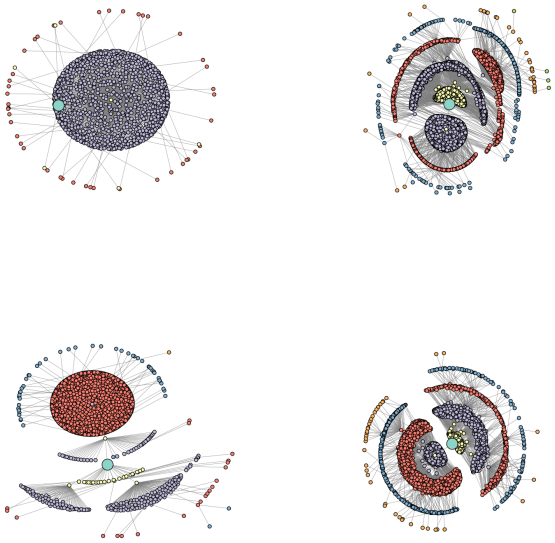

Fig. 7. Structure of the four largest cascades observed, all of which occurred on Twitter. Colors indicate node depth, with the large, green points corresponding to seed nodes.

seed node. Thus, not only are large cascades infrequent in these online domains, but even when present they do not resemble the multi-step epidemics that arise in prevalent theoretical models of diffusion [Watts 2002; Kempe et al. 2003; Leskovec et al. 2007], in which cascades become large by spreading for many generations. Referring again to the 1918 influenza epidemic, although it infected more than 500 million people, individuals typically infected no more than a handful of others [Mills et al. 2002], and hence its scale was determined almost entirely by the large number of generations over which it spread. By contrast, Fig. 7 shows that the largest adoption cascades we find spread for at most a few generations before dying out. In two of the top four cases, in fact, almost all adoptions are directly attributable to a single user with a very high follower count. Moreover, the first and fourth largest cascades correspond to independent introductions of the same newly released music video by a popular artist (Justin Bieber), and thus benefitted from intense media attention, marketing, and advertising. In short, the largest adoption events that we observe occur through mechanisms qualitatively distinct from those in strictly peer-to-peer models of contagion. In turn, this observation may motivate a more general class of diffusion models that follow classical influence studies [Katz and Lazarsfeld 1955] in explicitly differentiating media "actors" from ordinary individuals.

An alternative explanation for the occurrence of such popular products is that truly viral processes exhibit certain critical features that are unlikely to arise in the kind of social media domains that we consider here. For example, email viruses such as the Melissa Bug and Code Red [Smith 2004] spread over many generations to infect millions of computers on a global scale. Like biological epidemics, email viruses are capable of propagating themselves automatically, and without consent from the infected. In all the domains we investigate, however, both the adoption and transmission of products are deliberate decisions, suggesting that willful adoption and transmission may be

moderating factors of epidemic spread. Yet another possible explanation is suggested by the example of respondent-driven sampling (RDS) [Salganik and Heckathorn 2004; Goel and Salganik 2010]—a type of snowball sampling in which survey respondents enlist the next wave of participants—where recruitment chains routinely propagate for many generations, and usually terminate only as a result of deliberate action by the survey organizers. The success of this sampling methodology, however, depends critically on the provision of substantial financial incentives for recruiting [Malekinejad et al. 2008]. Thus, as with email viruses, the feature that drives the viral nature of RDS is unlikely to be present in typical online adoption processes, which lack direct financial incentives.

These alternative mechanisms notwithstanding, the possibility remains that viral spread among networks of individuals is occasionally responsible for large online adoption events, with potentially substantial consequences, and that we have simply not witnessed any such events in our data. For example, Liben-Nowell and Kleinberg [2008] study the propagation of an Internet chain letter that does in fact appear to have spread virally for hundreds of generations over more than ten years. While understanding the process underlying this seemingly rare phenomenon is certainly an important goal, that such viral diffusion may occur in some instances should be weighed against the very large number and diversity of observations that fail to exhibit any of the requisite properties.

We conclude by noting that in addition to their scientific and theoretical implications, our findings have practical relevance. Specifically, although the idea that social media can spread in the manner of biological epidemics is provocative, our findings indicate that strategies based on triggering "social epidemics" are likely unrealistic. Rather, we believe that marketers and others interested in efficiently diffusing information should work to harness and enhance the potentially valuable gains from each seed, where even incremental improvements in pass-along rates can lead to substantial returns on investment [Watts and Peretti 2007]. By reducing sharing costs and including various calls to action in word-of-mouth campaigns, reliable improvements in performance are likely possible [Aral and Walker 2011], even if truly viral diffusion of social media remains improbable.

ACKNOWLEDGMENTS

The authors are grateful to Erin Carlson, Connie Chan, and Jeremy Johnstone for assistance with the Yahoo! Kindness project; to Rushi Bhatt for the Yahoo! Voice data; and to Sergei Matusevych and Ayman Shamma for the Zync data.

REFERENCES

ADAR, E. AND ADAMIC, L. 2005. Tracking information epidemics in blogspace. In *IEEE/WIC/ACM International Conference on Web Intelligence*. IEEE Computer Society, Compiegne University of Technology, France.

AHO, A., HOPCROFT, J., AND ULLMAN, J. 1974. *Design & Analysis of Computer Algorithms*. Addison-Wesley.

ANDERSON, R. M. AND MAY, R. M. 1991. *Infectious Diseases of Humans*. Oxford University Press, Oxford.

ARAL, S., MUCHNIK, L., AND SUNDARARAJAN, A. 2009. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences 106,* 51, 21544–21549.

ARAL, S. AND WALKER, D. 2011. Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Management Science*.

BAKSHY, E., HOFMAN, J., MASON, W., AND WATTS, D. 2011. Everyone's an influencer:

quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*. Association of Computing Machinery, 65–74.

BAKSHY, E., KARRER, B., AND ADAMIC, L. 2009. Social influence and the diffusion of user-created content. In *Proceedings of the tenth ACM conference on Electronic commerce*. Association of Computing Machinery, 325–334.

BASS, F. M. 1969. A new product growth for model consumer durables. *Management Science 15,* 5, 215–227.

COLEMAN, J., KATZ, E., AND MENZEL, H. 1957. The diffusion of an innovation among physicians. *Sociometry 20,* 4, 253–270.

DEAN, J. AND GHEMAWAT, S. 2008. Mapreduce: Simplified data processing on large clusters. *Communications of the ACM 51,* 1, 107–113.

DODDS, P. AND WATTS, D. 2005. A generalized model of social and biological contagion. *Journal of Theoretical Biology 232,* 4, 587–604.

DOMINGOS, P., MIKA, P., GOLBECK, J., DING, L., FININ, T., JOSHI, A., NOWAK, A., AND VALLACHER, R. 2005. Social networks applied. *IEEE Intelligent Systems 20*, 80–93.

DOMINGOS, P. AND RICHARDSON, M. 2001. Mining the network value of customers. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 57–66.

FERGUSON, T. S. 1989. Who solved the secretary problem? *Statistical Science*, 282–289.

GLADWELL, M. 2000. *The tipping point: How little things can make a big difference*. Little, Brown and Company.

GOEL, S., MASON, W., AND WATTS, D. J. 2010. Real and perceived attitude agreement in social networks. *Journal of Personality and Social Psychology 99,* 4, 611–621.

GOEL, S. AND SALGANIK, M. 2010. Assessing respondent-driven sampling. *Proceedings of the National Academy of Sciences 107,* 15, 6743.

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

IYENGAR, R., VAN DEN BULTE, C., AND VALENTE, T. W. 2010. Opinion leadership and social contagion in new product diffusion. *Marketing Science*.

KATZ, E. AND LAZARSFELD, P. 1955. *Personal influence: the part played by people in the flow of mass communications.* Free Press.

KEMPE, D., KLEINBERG, J., AND TARDOS, E. 2003. Maximizing the spread of influence through a social network. In *9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association of Computing Machinery.

KITSAK, M., GALLOS, L., HAVLIN, S., LILJEROS, F., MUCHNIK, L., STANLEY, H., AND MAKSE, H. 2010. Identification of influential spreaders in complex networks . *Nature Physics* 11, 888–893.

LE BON, G. 1896. *The Crowd: A Study of the Popular Mind*. Macmillan.

LESKOVEC, J., ADAMIC, L., AND HUBERMAN, B. 2007. The dynamics of viral marketing. *ACM Transactions on the Web 1,* 1, 5.

LESKOVEC, J., SINGH, A., AND KLEINBERG, J. 2006. Patterns of influence in a recommendation network. *Advances in Knowledge Discovery and Data Mining*, 380–389.

LIBEN-NOWELL, D. AND KLEINBERG, J. 2008. Tracing information flow on a global scale using internet chain-letter data. *Proceedings of the National Academy of Sciences 105,* 12, 4633.

LIU, Y., SHAFTON, P., SHAMMA, D., AND YANG, J. 2007. Zync: the design of synchronized video sharing. In *Proceedings of the 2007 conference on Designing for User eXperiences*. ACM, 1–8.

LOPEZ-PINTADO, D. AND WATTS, D. 2008. Social influence, binary decisions and col-

lective dynamics. *Rationality and Society 20,* 4, 399–443.
MACKAY, C. 1841. *Extraordinary popular delusions and the madness of crowds*. Richard Bentley.
MALEKINEJAD, M., JOHNSTON, L., KENDALL, C., KERR, L., RIFKIN, M., AND RUTHERFORD, G. 2008. Using respondent-driven sampling methodology for hiv biological and behavioral surveillance in international settings: a systematic review. *AIDS and Behavior 12*, 105–130.
MILLS, C., ROBINS, J., AND LIPSITCH, M. 2002. Transmissibility of 1918 pandemic influenza. *J. Am. Med. Assoc 287*, 2236–2252.
MOORE, C. AND NEWMAN, M. E. J. 2000. Epidemics and percolation in small-world networks. *Physical Review E 61,* 5, 5678–5682.
ROGERS, E. 1962. *Diffusion of innovations*. Free Press.
SALGANIK, M. J. AND HECKATHORN, D. D. 2004. Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological Methodology 34*, 193–239.
SHALIZI, C. R. AND THOMAS, A. C. 2011. Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods and Research 40*, 211–239.
SHAMMA, D. AND LIU, Y. 2009. Zync with me: Synchronized sharing of video through instant messaging. *Social Interactive Television: Immersive Shared Experiences and Perspectives*.
SMITH, G. S. 2004. Recognizing and preparing loss estimates from cyber-attacks. *Information Security Journal: A Global Perspective 12,* 6, 46–57.
SUN, E., ROSENN, I., MARLOW, C., AND LENTO, T. 2009. Gesundheit! modeling contagion through facebook news feed. In *Proc. of International AAAI Conference on Weblogs and Social Media*.
TAUBENBERGER, J. AND MORENS, D. 2006. 1918 influenza: the mother of all pandemics. *Rev Biomed 17*, 69–79.
VAN DEN BULTE, C. AND LILIEN, G. 2001. Medical innovation revisited: Social contagion versus marketing effort. *American Journal of Sociology*, 1409–1435.
WATTS, D. J. 2002. A simple model of information cascades on random networks. *Proceedings of the National Academy of Science, U.S.A. 99*, 5766–5771.
WATTS, D. J. AND PERETTI, J. 2007. Viral marketing for the real world. *Harvard Business Review May*, 22–23.
YOUNG, H. P. 2009. Innovation diffusion in heterogeneous populations: Contagion, social influence, and social learning. *American Economic Review 99,* 5, 1899–1924.