On Word-of-Mouth Based Discovery of the Web

Tiago Rodrigues
Universidade Federal de Minas Gerais, UFMG
Belo Horizonte, Brazil
tiagorm@dcc.ufmg.br

Fabrício Benevenuto
Universidade Federal de Ouro Preto, UFOP
Ouro Preto, Brazil
fabricio@iceb.ufop.br

Meeyoung Cha
Graduate School of Culture Technology, KAIST
Daejeon, Korea
meeyoungcha@kaist.edu

Krishna P. Gummadi
Max Planck Institute for Software Systems, MPI-SWS
Saarbrucken, Germany
gummadi@mpi-sws.org

Virgílio Almeida
Universidade Federal de Minas Gerais, UFMG
Belo Horizonte, Brazil
virgilio@dcc.ufmg.br

ABSTRACT

Traditionally, users have discovered information on the Web by browsing or searching. Recently, *word-of-mouth* has emerged as a popular way of discovering the Web, particularly on social networking sites like Facebook and Twitter. On these sites, users discover Web content by following URLs posted by their friends. Such word-of-mouth based content discovery has become a major driver of traffic to many Web sites today. To better understand this popular phenomenon, in this paper we present a detailed analysis of word-of-mouth exchange of URLs among Twitter users. Among our key findings, we show that Twitter yields propagation trees that are wider than they are deep. Our analysis on the geolocation of users indicates that users who are geographically close together are more likely to share the same URL.

Categories and Subject Descriptors

J.4. [**Computer Applications**]: Social and behavioral sciences Miscellaneous; H.3.5 [**Online Information Services**]: Web-based services

General Terms

Human Factors, Measurement

Keywords

Word-of-mouth, information diffusion, social networks, web content discovery

IMC'11, November 2–4, 2011, Berlin, Germany.

1. INTRODUCTION

Recently online social networking sites like Facebook and Twitter have emerged as a popular way of discovering information on the World Wide Web. In contrast to traditional methods of content discovery such as browsing or searching, content sharing in social networking sites occurs through *word-of-mouth*, where content spreads via conversations between users. For instance, users share links to content on the web with personal recommendations like "This is a must-see video."

While such word-of-mouth based content discovery existed long before in the form of emails and web forums, online social networks (OSNs) have made this phenomenon extremely popular and globally reaching. In fact, today social networking sites are known to be a major driver of traffic to many web sites [40]. For certain web sites, Facebook and Twitter drive, respectively, 44% and 29% of the traffic [35]. These OSNs are sharing tens of millions of web links every day, [41] and we expect that the amount of information exchanged by word-of-mouth in OSNs will grow over time.

In this paper, we present a detailed analysis of the word-of-mouth based content discovery on the web, by analyzing the web links (URLs) shared on a popular social platform, Twitter. We used the Twitter data in [14], which comprises 54 million user profiles, 1.9 billion follow links, and all 1.7 billion tweets posted by Twitter users between March 2006 and September 2009. Twitter is an ideal medium to study word-of-mouth based discovery of web content for several reasons. First, the core functionality provided by Twitter, *tweeting*, is centered around the idea of spreading information by word-of-mouth. Second, Twitter provides additional mechanisms like *retweet* (act of forwarding other people's tweet), which enable users to propagate information across multiple hops in the network through word-of-mouth. Third, thanks to URL shortening services, sharing URLs has become a common practice in Twitter. In fact, nearly a quarter of all tweets in our data contained URLs, and a total of 208 million URLs were shared during the three year period.

In order to study how the 54 million users in Twitter collaboratively discovered and spread web links, we built an information propagation tree for every URL that was shared during a random week in 2009. We considered both explicit (e.g., retweets) and implicit information flows (e.g., when a user shares a URL that has already been posted by one of

the contacts she follows, without citing the original tweet). By accounting for implicit information flows, our methodology produces much denser trees than previous work that considers only explicit links [27,44]—the number of edges in a tree has increased by a factor of 23.9.

Based on the propagation trees of URLs, we try to answer several key questions that are fundamental to understanding word-of-mouth based web discovery. The questions we ask include: Can word-of-mouth reach a wide audience? What kinds of content are popular in social media as opposed to the web in general? Does word-of-mouth give all content, including those published by unpopular domains, a chance to spread? What are the typical structures of word-of-mouth propagation trees? We discuss the implications of our findings for the design of word-of-mouth based marketing strategies and the role of social media.

The key findings of our work can be summarized as follows:

1. Word-of-mouth can be used to spread a single URL to a large portion of the user population, in some cases even to an audience of several million users.

2. Popular URLs spread through multiple disjoint propagation trees, with each of the propagation trees involving a large number of nodes.

3. Domains whose URLs or content is spread widely by word-of-mouth tend to be different from the domains that are popular on the general Web, where content is found primarily through browsing or searching. Our analysis shows that word-of-mouth gives all content—including those published on relatively unpopular domains—a chance to be propagated to a large audience.

4. Word-of-mouth in Twitter yields propagation trees that are wider than they are deep. Our finding is in sharp contrast to the narrow and deep trees found in Internet chain letters.

5. Our analysis of the geo-location of users reveals a significant correlation between propagation and physical proximity. Moreover, content tends to spread for short distances only on the first hops away from the content creator.

The rest of the paper is organized as follows. Section 2 describes the data and our methodology for constructing information propagation trees. In Section 3 we examine what kinds of content is popular in word-of-mouth discovery of the Web. In Section 4 we characterize several aspects of word-of-mouth propagation. Section 5 presents an study of the geo-location of users and how far word-of-mouth content travels around the globe. We discuss implications of findings in Section 6 and present related work in Section 7. Finally, we conclude in Section 8.

2. METHODOLOGY

Twitter is a prime example of an OSN where users discover Web content through word-of-mouth. In this paper, we used the Twitter dataset gathered in [14] and study the properties of word-of-mouth based Web discovery.

2.1 The Twitter Dataset

Data collection utilized the official Application Programming Interface (API) of Twitter and took over a month using 58 servers in Germany [14]. We had these servers whitelisted by Twitter so that they can send API requests rapidly. The data comprises the following three types of information: profiles of 54,981,152 users, 1,963,263,821 directed follower links among these users, and all 1,755,925,520 public tweets that were ever posted by the collected users. The oldest tweet in our dataset is from March 2006, when the Twitter service was publicly launched. The dataset does not include any tweet information about a user who had set his account private (8% of all users). Our dataset is near-complete because user IDs were sequentially queried from all possible ranges (0–80 million) at the time of data collection in September 2009. Therefore, it provides a unique opportunity to examine the largest word-of-mouth based URL propagation event in Twitter.

A Twitter user might follow another user to receive his tweets, forming a social network of interest. The node in-degree and out-degree distributions measured on this network are heavy-tailed, and the network topology is similar to those of other OSNs like Facebook. They can be fit well with a Power-Law distribution with exponents 2.19 for in-degree and 2.57 for out-degree (R^2=0.05–0.09%). While a very small fraction of users have an extremely large number of neighbors, the majority of users have only a few neighbors; 99% of users have no more than 20 in- or out-degree neighbors. The most popular users include public figures like Barack Obama, celebrities like Oprah Winfrey, as well as media sources like BBC. A social link in Twitter is directional. Unlike other OSNs, the Twitter network exhibits extremely low reciprocity; only 23% of all links are bidirectional, which means that high in-degree nodes are not necessarily high out-degree nodes.

2.2 URLs in tweets

We treat a URL as a clean piece of information that spreads in Twitter. The number of tweets containing URLs has increased rapidly over the years, as shown in Figure 1. Since 2009, on average 22.5% of tweets contain URLs, and as of September of 2009 more than 30% of tweets contain URLs. This is equivalent to sharing 1.3 million distinct URLs per day in 2009. The URL usage is even higher in retweeets: 47%. Interestingly, the number of retweets grew abruptly after July of 2008. This is because retweeting became a convention between users around this time [13].

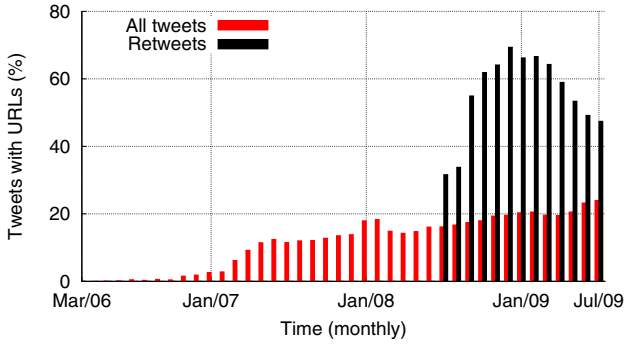


Figure 1: Usage of URLs on tweets over time

	Period	Distinct URLs	Tweets	Retweets	Users
Dataset 1	Jan 1, 2009 – Jan 7, 2009	1,239,445	6,028,030	295,665	995,311
Dataset 2	Apr 1, 2009 – Apr 7,2009	4,628,095	17,381,969	1,178,244	2,040,932

Table 1: Statistics of the two Twitter datasets analyzed

In analyzing the web links within tweets we found that the majority of the URLs (nearly 75%) were from URL shortening services (e.g., *bit.ly*), which substantially shorten the length of any web link. Hence we had to take into account several possible confounding factors such as when multiple short URLs refer to the same long URL or when the short URLs are recycled after not being used for some time (e.g., when there is no human visitor for 120 days, some URL shortening services allow a short URL to point to a new content). Therefore, we picked a large pool of URLs over several short time periods (so that these URLs refer to identical content), then resolved all the short URLs to the long URLs for data analysis in this paper.

Table 1 displays the summary of datasets we analyzed. Each week period contains several million distinct URLs. Because the samples are from a one week period, certain URLs were already in the process of word-of-mouth propagation. Hence, we scanned the entire tweet dataset to find all tweets that contain the URLs in Table 1 and additionally considered them in our analysis. We made sure that none of the added URLs were recycled. Due to space constraints, in the rest of the paper, we present results only for Dataset2 in Table 1. All the conclusions hold for Dataset1 as well.

2.3 Modeling information cascades

Below, we describe the model of URL propagation.

2.3.1 *Hierarchical tree model*

We build information propagation paths based on Krackhardt's hierarchical tree model [25]. A hierarchical tree is a directed graph where all nodes are connected and all but one node, namely the *root*, have in-degree of one. This means that all nodes in the graph (except for the root) have a single parent. Hence, an edge from node A to node B is added to the tree only when B is not already a part of the tree. An edge from node A to node B means that a piece of information was passed from A to B. While each hierarchical tree has a single root, there may be multiple users who independently share the same URL. In this case, the propagation pattern of a single URL will contain *multiple* trees and form a *forest*.

While we assume there is only a single parent for any intermediate nodes, in real life there may be more than one source who passed the same piece of information to a given user. In order to resolve the tie, we resort to the pattern seen in explicit links in our data. We found that, when users have multiple sources, more than 80% of them cite their last source. We attribute this pattern to the timeline interface of Twitter, which work as a stream, showing the last 200 tweets to the user, chronologically ordered. Hence, we assumed that each user received a URL from the most recent source. Our model is different from that proposed by Sun *et al.* [38], where they considered all friends who joined the same group on Facebook within the last 24 hours as valid sources.

Following the definitions proposed in [42], we call the users in the root of a hierarchical tree **initiators**. These users are the ones who independently shared URLs. We call all other nodes who participated in URL propagation **spreaders**. Initiators and spreaders make up the hierarchical tree. We call users who simply received a URL but did not forward it to others **receivers**. Later when we refer to the hierarchical tree structure, we do not include these users. For convenience, we collectively call all three types of users who potentially read the URL as its **audience**. Figure 2 depicts this relationship.

Figure 2: Terminology

2.3.2 *Explicit vs implicit links*

There are two major ways to reconstruct information propagation paths. One approach is to rely solely on the retweet convention. A retweet typically contains the original tweet content by someone else along with a "RT @username" or "via @username" text that cites the person who posted the tweet. Citation sometimes extends to the intermediary persons who also retweeted the same message. Identifying information propagation based on explicitly retweeted links has been a popular approach in previous work [27, 44]. Another approach, which we take in this paper, is to consider implicit links as well. As opposed to the former approach, implicit URL propagation occurs when a user, without necessarily citing the information source, shares a URL that has already been shared by one of his followees. In other words, an implicit URL propagation occurs when a user A publishes a URL after receiving it from a user B, but user A does not explicitly cite user B.

Since retweeting was not a built-in mechanism in Twitter in 2009 but was a convention that people started adopting over time, some Twitter users did not use the retweeting convention and even when they did, there were many forms of citation (e.g., via, retweeting, RT). Thus, ignoring these implicit links resulted in a sparsely connected graph as shown in Figure 3, which is based on a real example from our tweet data. In fact, we found that considering implicit links (as opposed to using only explicit retweet links) increases the number of edges in the information propagation paths by a factor of 23.9.

In order to verify the impact of using implicit links, we did some experiments varying the way in which the implicit link is chosen. Some examples include getting a random source (when the user received the URL from multiple sources) and considering an implicit link as valid only if it occurs within a specific time interval as valid (for example, if the retweet occurs within 24 hours of the tweet). In all the experiments we found similar results as the ones presented in this paper. As

more than 80% of the users usually retweet (explicity link) their most recent source, we decided to consider the most recent source in implicit links. Furthermore, as some retweets happen after a long time interval (for example, more than a month after the original tweet), we did not use any time interval limit.

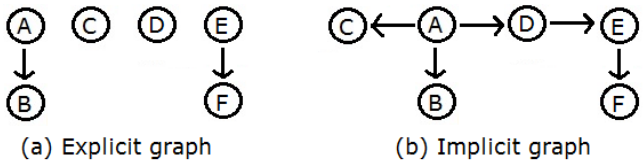


Figure 3: Explicit vs implicit graphs

3. WHAT CONTENT IS POPULAR

In this section, we examine which URL shortening services are widely used in Twitter, which web domains these short URLs point to, and what kinds of content is popular in word-of-mouth discovery of the Web.

3.1 URL shortening services

URL shortening services make it easy for Internet users to share web addresses by providing a short equivalent [9]. For example, a web link `http://topics.nytimes.com/top/news/business/companies/twitter/` can be shortened to `http://nyti.ms/1VKbrC` by a commercial service Bit.ly [2], which will redirect any request access to the original NYTimes website. URL shortening services allow otherwise long web addresses to be referred to in various OSNs like Twitter that often impose character limit in tweets and comments. There are hundreds of commercial URL shortening services. Hence, the same web address can have several short alternatives in services like tinyURL [6] and Ow.ly [4].

In order to identify whether a given web address is a short or long URL, we took a heuristic approach. We wrote a Python script to resolve a URL in a tweet by sending a web access request to that URL and comparing the domain of the original URL with that of the resolved URL. If the two domain names were different, we considered the URL in the tweet to be a short URL, otherwise, we considered it a long URL.

A total of 30 URL shortening services were in use from 2006 to summer 2009 in Twitter. Table 2 displays the top 10 services and their share of tweets. Usage of the top two services tinyurl.com and bit.ly make up more than 90% of the total usage. Between January to April of 2009, we find that bit.ly doubled its presence. The 3rd ranked service is.gd also continued to gain presence in Twitter.

3.2 Popularly linked web domains

We next checked whether URLs popularly shared on Twitter come from major web domains in the Internet (such as nytimes.com or google.com). Our motivation is to verify a widely held belief that word-of-mouth can help popularize niche or esoteric information from domains that are not otherwise very popular. We used the translated long URLs for this analysis and the rest of the paper, and grouped the URLs based on their domain names.

Rank	Web Domain	Dataset2	Dataset1
1	tinyurl.com	4,398,940 (68.2%)	1,883,032 (81.4%)
2	bit.ly	1,530,613 (23.7%)	262,171 (11.3%)
3	is.gd	493,124 (7.6%)	142,497 (6.2%)
4	snipurl.com	27,146 (0.4%)	24,606 (1.1%)
5	hugeurl.com	1,578 (0.0%)	841 (0.0%)
6	url.ca	1,116 (0.0%)	451 (0.0%)
7	xrl.in	361 (0.0%)	250 (0.0%)
8	u.nu	282 (0.0%)	139 (0.0%)
9	simurl.com	216 (0.0%)	6 (0.0%)
10	doiop.com	101 (0.0%)	90 (0.0%)

Table 2: Top 10 URL shortening services in 2009

In total, there were 4,638,095 long URLs that came from 429,551 distinct web domains. The top 20% of the web domains accounted for 95% of these URLs. We ranked the domains based on the number of distinct URLs that belong to the domain as well as the total size of the audience reached by URLs belonging to the domain. Experiments using both ranking methods had similar results. We compared the list of top domains in the resulting rankings with the list of top ranked domains in the general Web published by Alexa [1]. Table 3 displays the top-5 domains based on the number of URLs, their description, the fraction of all URLs that belong to the domain, and their rank from Alexa.com.

Rank	Top list	Description	URLs	Alexa rank
1	twitpic.com	photo sharing	8.5%	103
2	blip.fm	music sharing	3.0%	6,736
3	youtube.com	video sharing	2.1%	3
4	plurk.com	social journal	2.1%	1,146
5	tumblr.com	blog	1.4%	100

Table 3: Top-5 domains in Twitter (April, 2009)

The most popular domain, twitpic.com, accounted for 8.5% of all URLs in the tweet data. The coverage of the other top domains quickly drops with decreasing rank. The Alexa ranking shows that the top-5 domains are quite different from the top list in the Web. Only youtube.com is within the top 10 sites from alexa.com. The top-5 list in Alexa includes major portals and search engines (Google) and portals (Yahoo, Live). We also found that Twitter users often share user-generated content, as seen in the table. Twitter users share photos (twitpic.com), videos (youtube.com), blog articles (techcrunch.com), as well as participate and promote social events (abolishslavery.com and earthday.net).

Figure 4 shows the fraction of top-K domains that also appear in the top-K domain list from Alexa.com for various values of K. The bar plot also shows a comparison to the top URLs identified by the size of the audience reached within Twitter (including initiators, spreaders, and receivers. The overlap is minimal; fewer than 30% of the top-K domains in Twitter overlap with that of the general Web, for all ranges of K=100,$\cdots$,1000. This finding suggests that as word-of-mouth becomes a dominant source of discovering information, a different set of domains might become popular in the Web in the near future.

A recent work characterizing the usage of short URLs on Twitter also presented an analysis comparing the popularity of domains on Twitter and on the general Web [9]. Although the authors of that work considered only the domains pointed by short URLs, which differs from our analysis, they also found that the most popular domains shared on Twitter differs significantly from the general Web case.

Rank	URL domain	Audience	Description
1	wefollow.com	28M	Social application that suggests list of users to follow
2	facebook.com	14M	Social network (warning page)
3	abolishslavery.org	4.5M	Social organization dedicated to combating human traffic (initial page)
4	twitpic.com	4.5M	Photo sharing (photo published by famous actor Ashton Kutcher)
5	youtube.com	4.5M	Video sharing (popular comedy video with title "David After Dentist")
6	tweetvalue.com	4.3M	Application that measures the value of a Twitter account (initial page)
7	techcrunch.com	3.6M	Blog (article with rumors about Google in talks to buy Twitter)
8	earthday.net	3.3M	Social organization dedicated to the Earth's natural environment (initial page)
9	twibes.com	3M	Application to find people with similar interests on Twitter (initial page)
10	latenightwithjimmyfallon.com	2.4M	TV Show from NBC (initial page)

Table 4: Top 10 URLs domains in terms of the audience size reached by the most popular URL

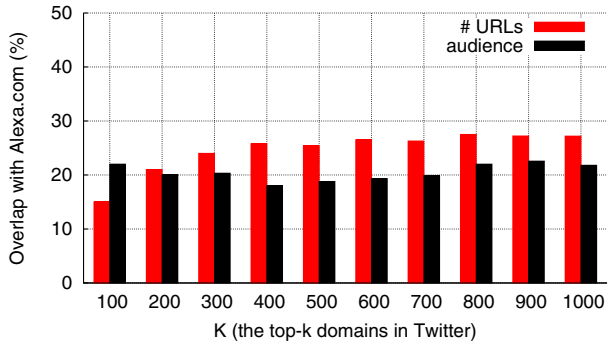


Figure 4: Popular domains in Twitter and Alexa rank

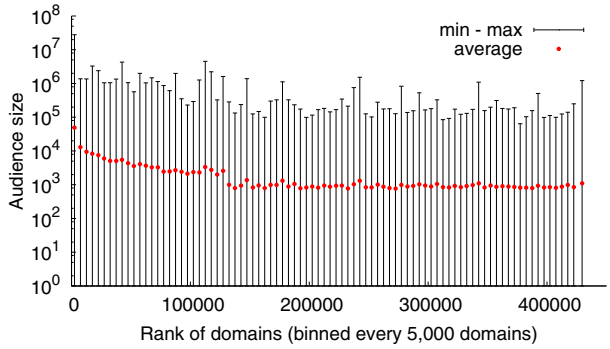


Figure 5: Audience size across the domain ranks

3.3 Popular individual URLs

Next we focus on popularity of individual URLs within domains. Of particular interest to us is the hypothesis that word-of-mouth gives all URLs and content a chance to become popular, independent of popularity of the domain it comes from. The hypothesis is rooted in the observation that anyone could identify an interesting URL and start a viral propagation of the URL, independent of the reputation or popularity of the domain where it is published.

To verify this hypothesis, we computed the size of the audience reached by individual URLs within each domain. Figure 5 plots the minimum, the average, and the maximum size of the audience for individual URLs within each domain. Given the large number of domains (over 400,000 of them) we ranked the domains based on number of URLs, grouped them into bins of 5,000 consecutively ranked domains and plotted one data point for each bin. It is striking to observe that URLs from some unpopular domains beyond the rank of 300,000 reached an audience that is comparable to the size of the audience reached by URLs from the most popular domains. On average, URLs in the top 5,000 domains reached 49,053 users, while URLs from the bottom 5,000 domains reached 1,107 users. Although URLs from top domains reached a 44 times larger audience than those from bottom domains, there do exist individual URLs from bottom domains that reach as large as audience as the most popular URL from the top domains.

Thus, word-of-mouth does offer a chance for all content to become popular, independent of the domain it is published in. Previous work on book and DVD recommendations [29] showed that viral marketing is effective for niche products compared to mass marketing. Our analysis suggests a similar trend.

3.4 Content types

With millions of URLs published per day on Twitter, several different types of contents are shared. A natural question that arises from this observation is whether the type of content affects the word-of-mouth propagation dynamics.

In order to identify the content type, we first select several interest categories based on Open Directory Project (DMOZ), which is a human-edited directory of the Web [3]. DMOZ's content classification relies on the fact that a domain name has a hierarchical structure. So, a domain name of a web address could potentially be used to identify the specific content category, given the list of predefined classifications for many web domains as in the DMOZ service. For example, nytimes.com is classified in the news category; last.fm is classified as the music category.

Among various categories DMOZ supports, we picked 5 categories of interest: photo sharing, videos, music, news, and applications, and downloaded the list of web domains in each of these categories. In total, the DMOZ listing contained 343 domains across all five categories. For the application category, we used the list of applications that are widely used within Twitter, such as tweetdeck.com and wefollow.com.[1]

Table 5 displays the share of each topical category in Dataset2. Matching the domains of the URLs in Dataset2 with the domains listed in these five categories in DMOZ, we were able to successfully categorize 17.6% of websites, although a much larger fraction of users (43.7%) were cov-

[1] `http://www.dmoz.org/Computers/Internet/On_the_Web/Online_Communities/Social_Networking/Twitter/`

ered. The most popular category is photo sharing, which has 458,662 URLs, posted by 271,138 users on 735,137 tweets. The average audience reached by each photo's URL is 436 users. The second most popular category is music, followed by videos, news and applications.

While photo sharing seems a dominant activity in Twitter as opposed to news or application sharing, the set of most popular individual URLs are from a diverse set of topical categories as shown in Table 4. The most popular URL was the social application, wefollow.com, which reached an audience of 28 million (i.e., nearly half of the entire Twitter network).

Although our analysis the different content types shared on Twitter is an interesting aspect of our study, we note that our methodology has some limitations. First, we were able to categorize only 17.6% of the URLs, which might not be representative. Second, DMOZ is often criticized for its lacks of representativeness and transparency [7], but it is not easy to categorize content to begin with, and we manually checked the list of web domains in each category we used from DMOZ. Moreover, our main interest in the URL categorization is at understanding the similarities and differences between the propagation of different content types on Twitter. We are not trying to say that Twitter users share more photos than videos or news, for example.

4. THE SHAPE OF WORD-OF-MOUTH

This section presents an analysis of the size and shape of word-of-mouth based URL propagation patterns in Twitter.

4.1 How large is the largest word-of-mouth?

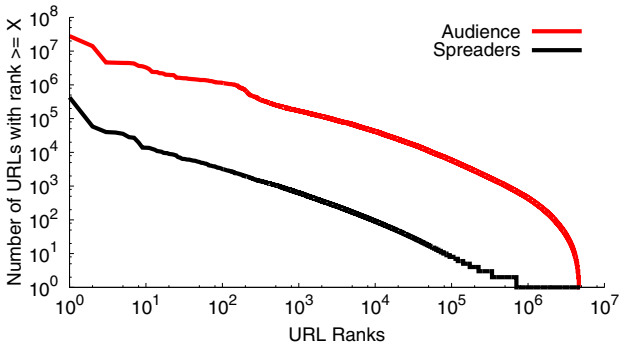
We examine the skew in popularity across different URLs. Figure 6(a) shows the size distribution of spreaders and audience for URL propagations in Twitter. An average URL was spread by three users and gained an audience of 843 users through word-of-mouth. In contrast, the most popular URL engaged 426,820 spreaders and reached an audience of 28 million users, which is more than half of the entire Twitter network. The power of word-of-mouth extends beyond the few most popular URLs. Each of the 100 most popular URLs reached an audience of more than 1 million users and 15% of the URLs reached an audience of over 1000 users.

The difference between the number of spreaders and the size of the audience is nearly two orders of magnitude, for popular URLs as well as niche URLs that have only a few spreaders. This demonstrates the potential of word-of-mouth in reaching a large audience. As opposed to a typical web page that is viewed by individual visitors, content shared in word-of-mouth fashion is *collaboratively* shared by other visitors who liked it and can reach a much larger audience.

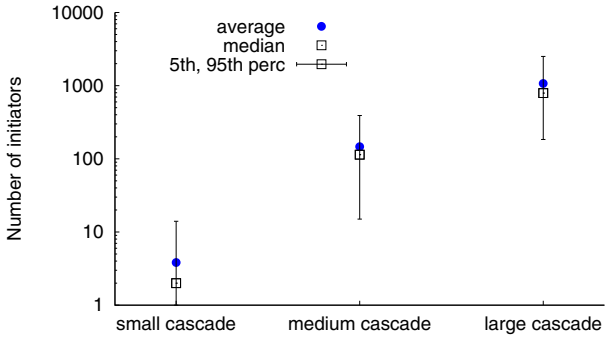
Both of the distributions for spreaders and audience exhibit power-law behavior (a straight line waist in a log-log plot). The best fit power-law exponents of these distributions $y = cx^{-\alpha}$ were $\alpha = 1.71$ for spreaders, and $\alpha = 1.98$ for audience, indicating that the skew in popularity among the most popular and the least popular URL became slightly more severe due to audience.

4.2 The role of initiators

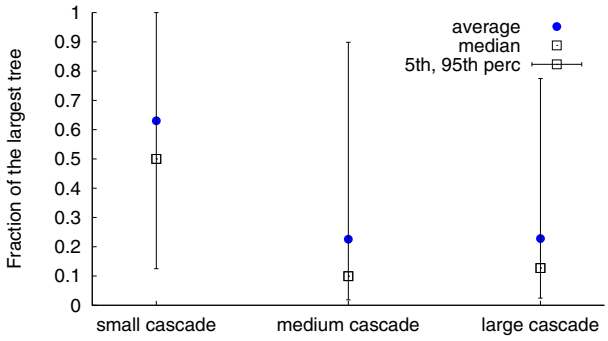
We classified users into three types based on their position within a cascade: initiators, spreaders, and receivers (Figure 2). Initiators are at the root of each cascade tree and share URLs to others independently in Twitter. In this subsection, we investigate the role of initiators in particular and ask the following questions: To what extent can initiators alone reach a large audience (without the help of spreaders)? How many initiators share the same URL? Is having multiple initiators essential for yielding a large cascade?



(a) The size of word-of-mouth



(b) The number of initiators per URL



(c) Roles of initiators by the largest subtree

Figure 6: Characteristic shape of word-of-mouth information propagation in Twitter

In Twitter, nearly 90% of all URLs are introduced only by initiators without involving any spreaders. URLs that were propagated further by spreaders went multiple hops in the Twitter social graph and gained a 3.5 times larger audience than those spread by initiators. This implies that while initiators' role is dominant and that most URLs propagate only 1-hop in the network, multi-hop propagation by spreaders can extend the readership of URLs by a significant amount.

Next we investigated the relationship between the num-

Category	URLs	Users	Tweets	Audience
Photos	458,662 (9.7%)	271,138 (13.3%)	735,137 (4.0%)	436
Music	181,676 (3.8%)	46,483 (2.3%)	338,654 (1.8%)	316
Videos	123,412 (2.6%)	261,081 (12.8%)	509,975 (2.8%)	877
News	58,467 (1.2%)	113,667 (5.6%)	305,911 (1.7%)	2,492
Applications	15,223 (0.3%)	198,499 (9.7%)	370,047 (2.0%)	3,285

Table 5: Summary information of categories of URLs

ber of initiators and cascade size. In order to focus on URLs with multiple initiators, we focus on URLs with at least two participants (i.e., initiators or spreaders). We group URL cascades into three types: small (1,10), medium [100,1000), and large [1000,∞). Figure 6(b) shows the number of initiators per URL for the three types of cascades. The plot shows the 5th percentile, median, average, and 95th percentile values. The number of initiators are orders of magnitude larger over the cascade size. The median number of initiators is 2, 114, and 792 for small, medium, and large cascades, respectively. Furthermore, a very few URLs (0.2%) had more than 100 initiators who independently shared the same URL. While certain large cascades only involved a single initiator, the plot indicates that the number of initiators does affect the size of the cascade for most URLs—larger cascades likely involve more users who independently share the URL.

While there seems to be a relationship between the number of initiators and cascade size in general, it is not clear whether all of the initiators contributed equally in obtaining audience, or whether a few major initiators played a significant role. Figure 6(c) investigates the fraction of the largest subtree for each URL over cascade size. The plot shows much variability, shown by the wider range of the 5th and 95th percentiles. Nonetheless, the median and the average data points shows a clear trend.

For small cascades, the fraction of total audience reached in the largest subtree (i.e., by a single major initiator) is nearly 50%. For larger cascades, however, the fraction of the largest subtree is marginal (10-20%). This implies that a single initiator's role is likely limited in these cases and that popular content usually propagates through several different propagation trees. In fact, we observed a strong positive correlation between the number of initiators and the total size of the audience (Pearson's correlation ρ=0.7171).

On Digg, a social news aggregator that allows users to submit links to and vote on news stories, a story of general interest usually spreads from many independent seed sites, while a story that is interesting to a narrow community usually spreads within that community only [28]. This observation might also be true in the case of URLs propagated in Twitter, but we leave the investigation for future works.

4.3 The shape of word-of-mouth

Having investigated the size of URL cascades, we next examine the overall shape of URL propagations. We define the following terminology. We refer to the maximum hop count from root to any of the leaf nodes as the *height* of the tree. We define the *width* of the tree as the maximum number of nodes that are located at any given height level. For instance, a two-node cascade graph has height of 2. The cascade graph in Figure 3(b) has a width of 3 and a height of 4 (i.e., the longest path being A-D-E-F). Because a single URL propagation may have multiple tree structures, we consider only the largest propagation tree for each URL and examine its width and height.

Figure 7(a) shows the distributions of height and width for all URLs. Nearly 0.1% of the trees had width larger than 20, while only 0.005% of the URLs had height larger than 20. This suggests that cascade trees in Twitter are wider than they are deep. In fact, the maximum observed width of any propagation tree was 38,418, while the maximum observed height was 147—-a difference of two orders of magnitude.

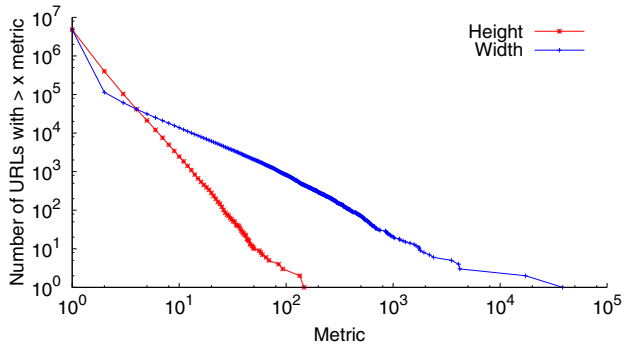
Figure 7(b) shows the relationship between the width and height level, as the 10th, 50th (median), and 90th percentile width over every height level. We grouped trees based on their size (according to the number of spreaders). Small trees with fewer than 100 spreaders tend to have a very narrow shape of width 1 or 2 throughout the height level. Larger trees with more than 100 spreaders were widest at low heights and the width decreased slowly towards the leaf nodes. Interestingly, the median width remained near 10 even at heights above 80. A visualization revealed that these large trees occasionally included bursts at all height levels, i.e., the branch out factor is suddenly large at particular spreaders. We found not one but multiple such bursts for every popular URL.

Finally, Figure 7(c) shows the size of a typical cascade for the five different types of URLs: photos, music, videos, news and applications. We find several interesting differences across content types. Videos propagations likely involve a larger cascade tree; more than 30% of videos URLs involved at least two participating users who shared the URL. News and applications propagation also involved multiple users (28% and 23%, respectively), while photos and music were mostly shared by a single initiator (90% and 97%, respectively). These observation indicates that the type of content affects the potential of the eventual cascade size. The probability of involving 10 or more users in spreading is around 10% for news, applications and videos, while it is only 1% for photos and music.

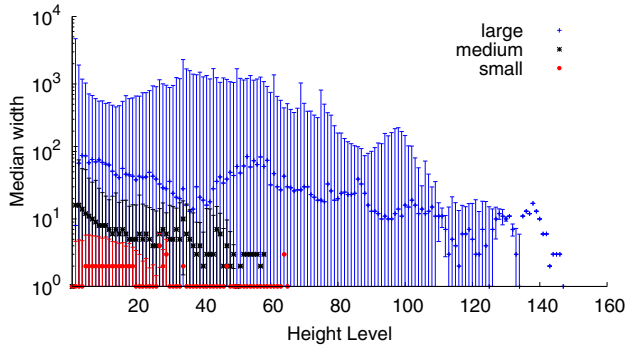
4.4 Comparison of cascade shapes

A large difference between the height and width in Figure 7(a) indicates that cascade trees are likely shallow and wide. This word-of-mouth propagation pattern is in stark contrast to other patterns of cascade.

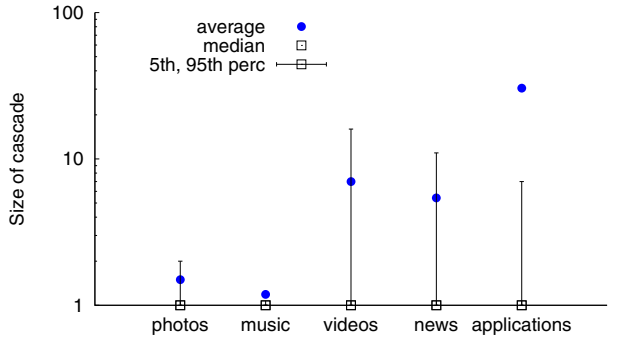
Recent work [31] demonstrated that massively circulated Internet chain letters proceed in a narrow but deep tree-like pattern, continuing for several hundred propagation hops. The Internet chain letter on the circulation of a petition on the US-Iraq war reached 18,119 users with a median height of 288 and the maximum width of 82. Table 6 shows a comparison between the height and width of Internet chain letters and the Twitter propagation graph. We estimated the size of smaller Internet chain letters following the model



(a) Height and width distributions



(b) Cascade size distributions



(c) Size per content type

Figure 7: Height and width of the largest cascade subtree

described in [5, 31]. In the table, we grouped Twitter propagation graphs of similar sizes (900-1100 nodes and 100-300 nodes) and showed the average height and width of those trees.

Twitter cascades are much wider than they are deep, unlike those of the Internet chain letters. We believe the broadcast design of Twitter directly influence such difference. As an example, consider that 5 friends (namely, A, B, C, D and E) want to exchange messages. On e-mail systems, the sender of each message needs to choose all receivers for that message. In our example, user A might send an e-mail to B and C, who decides to forward it to D and E. In this case, the tree is of height 3 and width 2. What happens on Twitter is exactly the opposite: the receivers choose all senders from whom they want to receive messages automatically. The senders just need to send the messages and a huge audience is reached quickly. In our simple example, if B, C, D and E follow user A, if A send a tweet all of them will receive it at once. In this case the tree is of height 2 and width 4. Another possible explanation for wider cascades on Twitter is the size of the messages. As Twitter allows only 140 characters, users can quickly read and forward a message. If people receive a long email, it is possible that a lot of recipients will not read and, consequently, not forward it. We leave in-depth investigation of the reasons for when wide cascades occur in Twitter as future work.

Data Source	Nodes	Height	Width
PNAS [31]	18,119	288 (med)	82
Twitter	26,227	23 (max)	17,255
PNAS (estimated)	980	16 (med)	4
Twitter (900-1,100)	980±0.02	20±0.25	398±0.27
PNAS (estimated)	162	3 (med)	1
Twitter (100-300)	162±0.02	10±0.03	86±0.03

Table 6: Shape comparison between Internet Chain letter and the Twitter propagation graph

5. CONTENT DISTRIBUTION

So far we investigated several key characteristics of word-of-mouth based URL propagation in Twitter. The observed propagation patterns have direct implications on systems, especially on content distribution and caching strategies. In this section, we revisit some of these observations and use geo-location information of users to examine how far word-of-mouth content travels around the globe.

For this, we need to know the location of users who post and receive tweets. The location information written in user profiles of Twitter is in free text form and often contains invalid location like "Mars" making it difficult to automate the process. We filtered out invalid locations and inferred plausible locations of users by using the Google Geocoding API [21], which converts addresses or city names written in free text form into geographic coordinates of latitude and longitude. In total, we identified the location of 1,096,804 users. In the remainder of this section, we only consider the network and the URL propagation patterns among these one million users.

5.1 Content producers and consumers

We first investigate physical proximity between content producers and consumers in Twitter. Here, a content producer represents a user who posts a URL independently of others (i.e., root nodes in any cascade tree) and a content consumer represents all other nodes in a cascade tree. Figure 8 shows the distribution of physical distance between any two users in the word-of-mouth URL propagation. Physical distance is computed based on the latitude and longitude information of two users. The distance is in units of 10km, representing a local community. The graph shows the probability distribution function for each distance d, which is the physical distance among all user pairs (u, v), where either (1) user u explicitly retweeted the URL that another user v shared or (2) user u follows another user v and shared the same URL after v posted it on Twitter. If either of these two conditions holds, we say that there is a *propagation* link from user u to user v.

For comparison purposes, Figure 8 also shows the distance distribution for two users who have a (unidirectional) follow link between them. We call this a *friendship* link. The friendship links represent the full potential of content distribution through word-of-mouth (i.e., when every follower actively reads or consumes the URL she receives from others).

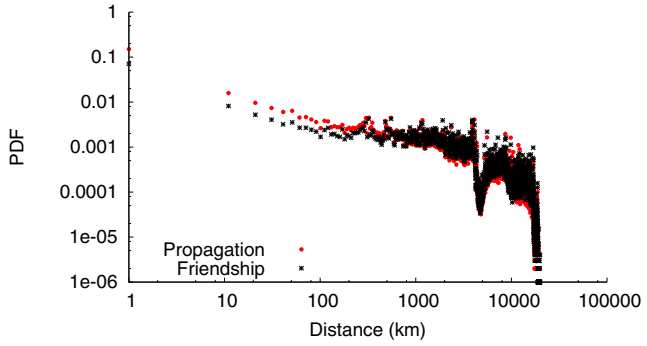


Figure 8: Physical distance of Twitter friendship links and URL propagation links

We observe a significant correlation between the content propagation probability and physical proximity of users. That is, users within a short geographical distance (e.g., 10 km) have a higher probability of posting the same URL than those users who are physically located farther apart. The current OSN infrastructures could exploit this physical proximity between content producers and consumers. Moreover, a significant correlation between the friendship and physical proximity is also observed. This is expected as users tend to interact more with other users who are physically nearby. Liben-Nowell *et al.* [32] previously found a strong correlation between friendship and geographic location among LiveJournal users.

Interestingly, the correlation between the content propagation probability and physical proximity of users is slightly higher than that observed for having a friendship link. This might be explained by the fact that Twitter users follow not only their friends, but also media companies and celebrities, as well as distant users that post content that is valuable to them. However, when it comes to retweeting other users' messages, Twitter users chose tweets posted by those geographically nearby.

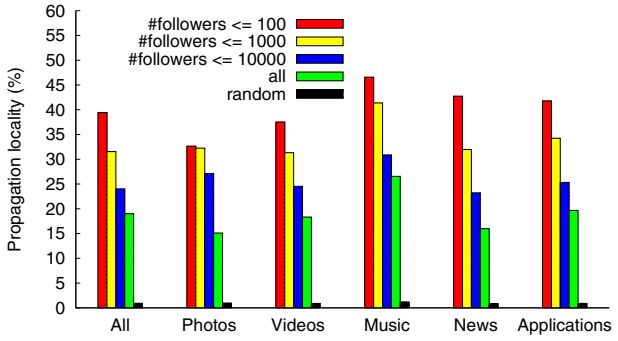


Figure 9: Distribution locality across content types

We tested if the level of locality changes according to the popularity of content producers and different content types. The first set of bars in Figure 9 shows the probability that content producers and consumers are located within a distance of 50km, which roughly represents a large metropolitan area. In order to separate out the impact of producer popularity, we grouped content producers into three groups based on their in-degree. Also, in order to see the representativeness of the results, we randomized the location between all users in our dataset by shuffling the location tags of users and computed the distance between them (shown as 'random' in the figure).

First, we focus on the impact of content producer's popularity on distribution locality and examine the first set of bars labeled 'All' content type in the x-axis. Overall, about 24% of the users who propagated content are physically close to very popular content producers who had more than 1000 followers, 32% are close to content producers with between 100 and 1000 followers, and 39% are close to content producers with less than 100 followers. This result indicates that producers with a small number of followers tend to incur content propagations to geographically nearby locations. On the other hand, content producers with a large number of followers tend to be celebrities or well-known people and incur content propagations across wider areas.

Next, we focus on the impact of content types on distribution locality. Figure 9 also shows the result for different types of content (photos, videos, music, news and applications). Interestingly, the locality for photos is the lowest, while the locality for music is the highest. News and applications had much stronger local appeal than photos and videos. Overall, a non-negligible fraction (15-25%) of content propagated locally. We do not know the reasons for why certain content type has more local appeal than the others. We leave investigating these reasons as exciting future work.

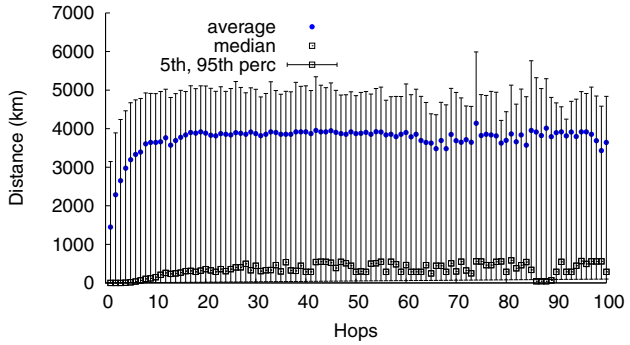


Figure 10: Distance between initiators and spreaders

Motivated by the overall high content locality at 1-hop users, we look at the distance a URL travels as it is further propagated by users within 2- and 3-hops away from the content creator. Figure 10 shows the geographical distance related to the content creator (initiator) as a function of the height level of spreaders on the hierarchical trees. Clearly, content tends to spread short distances on the first hops. As soon as friends-of-friends join the cascade and shares the same URL, it reaches users located in different regions and, consequently, reach distant locations.

The observations that social content produced by users

Rank	Country	Users	Tweets	Rank	U.S. State	Users	Tweets
1	United States	613,458 (55.9%)	6,319,585 (34.1%)	**1**	California	86,532 (19.7%)	910,379 (21.1%)
2	United Kingdom	86,613 (7.9%)	856,318 (4.6%)	**2**	Texas	36,509 (8.3%)	358,013 (8.3%)
3	Canada	54,021 (4.9%)	525,869 (2.8%)	**3**	Florida	24,389 (5.5%)	238,558 (5.5%)
4	Brazil	51,047 (4.7%)	395,433 (2.1%)	**4**	New York	21,191 (4.8%)	194,971 (4.5%)
5	Australia	31,963 (2.9%)	276,762 (1.5%)	**5**	Illinois	18,070 (4.1%)	167,597 (3.9%)
6	Germany	27,069 (2.5%)	283,733 (1.5%)	**6**	Washington	17,352 (3.9%)	165,495 (3.8%)
7	Iran	15,827 (1.4%)	149,887 (0.8%)	**7**	Georgia	16,842 (3.8%)	144,895 (3.4%)
8	Netherlands	15,812 (1.4%)	208,046 (1.1%)	**8**	Massachusetts	16,817 (3.8%)	159,464 (3.7%)
9	Japan	14,087 (1.3%)	222,825 (1.2%)	**9**	Pennsylvania	16,027 (3.6%)	142,265 (3.3%)
10	India	12,750 (1.2%)	140,124 (0.8%)	**10**	Ohio	13,064 (3.0%)	126,457 (2.9%)

Table 7: Summary information of geographical location of users

with a small number of followers is usually consumed by users that are located within a small physical distance of the content creator could be exploited for caching design and content delivery networks (CDNs).

5.2 Expected traffic pattern

Finally, we study the potential traffic pattern of content that is produced and consumed by users in the Twitter dataset. In Dataset2, 56% of the users are from the United States. In order to conduct a more detailed analysis of the traffic pattern, we inferred a finer grained geo-location information of these U.S. users. We used publicly available data from the U.S. National Atlas and the U.S. Geological Survey to map the inferred latitudes and longitudes into their respective county and state within the United States. In total, we were able to obtained county-level information for 440,036 users.

Table 7 displays the distribution of users and their tweets for the 10 most populated countries and the 10 most populated states in the U.S. based on our Twitter data. Users from United States are well distributed among the states, as the most popular, California, contain no more than 19.7% of all U.S. users.

We then estimated the volume of traffic generated from the U.S. through Twitter by focusing on the subset of URLs linking to photos and videos. In particular, we focused on two most popular services: twitpic.com and youtube.com. For all URLs linking to these services in Dataset2, we collected information about the exact size of each twitpic.com photo and the duration of each youtube.com video. We could not get the exact size of the videos through the YouTube API, hence we estimated the size of each video based on the play length and the video encoding rate of each video. Table 8 shows the statistics for the 10 states that generated the highest traffic volume. The traffic volume was calculated by multiplying the size of each photo or video with the number of tweets from users of each state who posted the URL of the content.

Although the traffic pattern shown in Table 8 is based on our estimation, we conjecture that the real traffic distribution would look similar because we conservatively chose active content consumers who actually looked at the content and decided to share it further. We observe similar distributions of traffic pattern for both photos and videos. The most popular state, California, is responsible for 24.2% and 23.9% of the traffic of photos and videos, respectively. While this number is slightly higher than the user population (19.7%), we find that content producers are located widely across different states. In fact, 9 out of the 10 most populated states appear in Table 8.

Considering that content in Twitter is typically produced by geographically-diverse users but consumed locally, one could allow users to upload content to a local server in the corresponding geographical area, like a server located in a city center, instead of a national server like in the current distribution architecture. Such a content distribution mechanism could significantly reduce the amount of wide-area bandwidth needed to upload the same URL to a centralized, remote server. Additionally, the trend towards local consumption and the fact that users are uniformly distributed across different locations indicates that placing a server in every region can reduce the amount of cross-region traffic.

5.3 Summary

Information dissemination among individuals located within a short physical proximity has been used to explain a number of phenomena in society, such as the proliferation of specific industries in a certain region [36] and individuals employment status [39]. Here, we were particularly interested in understanding how far and over what distance social content is generated and consumed. Given that OSNs use the infrastructure originally designed for web workloads to deliver social content, understanding geographical aspects of word-of-mouth propagation patterns can unveil potential opportunities for improving the current designing of OSN content delivery [34].

6. DISCUSSION

In this paper, we made several key observations. For two of these observations we would like to provide additional explanations. First, we showed that word-of-mouth could reach a large audience. Second we noted that the shape of the word-of-mouth propagation tree is more wide than deep.

• The power of word-of-mouth

In Section 3, we showed that URLs could reach a wide audience of several tens of million users through word-of-mouth. Most word-of-mouth events, however, involved a single user who shared URLs and reached only a small audience. This finding conforms to the observation by Watts [43] that global cascades are rare, but by definition are extremely large when they do occur.

In studies of information propagation, a great amount of attention has been devoted to the cascade size distribution and the underlying mechanism that shapes it. Unfortunately, due to a lack of empirical data detailing such cascade size distributions, synthetic distributions have been used to describe the phenomenon [43]. Largely, power-law and bimodal distributions imply the trend of large events occurring infrequently, although these distributions are quite different

Photos				Videos			
Rank	**U.S. State**	**Traffic (GB)**	**Tweets**	**Rank**	**U.S. State**	**Traffic (hours)**	**Tweets**
1	California	2,411 (24.2%)	32,395 (24.6%)	**1**	California	854 (23.9%)	11,813 (24.1%)
2	Texas	985 (9.9%)	13,481 (10.2%)	**2**	Texas	345 (9.6%)	4,549 (9.3%)
3	Florida	688 (6.9%)	8,951 (6.8%)	**3**	Florida	209 (5.8%)	2,755 (5.6%)
4	Washington	517 (5.2%)	6,190 (4.7%)	**4**	Illinois	178 (5.0%)	2,425 (4.9%)
5	Georgia	452 (4.5%)	5,473 (4.2%)	**5**	Washington	155 (4.3%)	2,093 (4.3%)
6	Illinois	411 (4.1%)	5,571 (4.2%)	**6**	Massachusetts	152 (4.3%)	2,223 (4.5%)
7	Pennsylvania	392 (3.9%)	5,340 (4.1%)	**7**	Pennsylvania	148 (4.1%)	1,978 (4.0%)
8	Ohio	338 (3.4%)	4,362 (3.3%)	**8**	Georgia	140 (3.9%)	2,023 (4.1%)
9	Massachusetts	282 (2.8%)	3,800 (2.9%)	**9**	Ohio	115 (3.2%)	1,631 (3.3%)
10	Arizona	275 (2.8%)	3,588 (2.7%)	**10**	Oregon	97 (2.7%)	1,316 (2.7%)

Table 8: Total volume of traffic potentially initiated by Twitter on each state (United States), for Twitpic photos and YouTube videos

and represent different underlying mechanisms. In this respect, the Twitter data provides a first opportunity to study size distributions.

The size distribution of word-of-mouth was best fit with the power-law distribution $y = cx^{-\alpha}$. In contrast to the exponents α of 2.19–2.56 seen in the in-degree and out-degree distributions of the topology, the exponents 1.71–2.23 observed in cascade sizes are smaller. This difference may be due to the collaborative act of sharing in Twitter. If each user were to share a URL without the help of any other user, then the size distribution for such a cascade should be the same as the in-degree distribution. However, when users collaboratively spread the same URL, the gap between the most popular user and the least popular user (based on in-degree of a user) becomes less important to the success of cascades, hence yielding a smaller exponent.

• The shape of propagation trees in Twitter

In Section 4, we analyzed the shape of propagation trees and found that word-of-mouth in Twitter yields trees that are much wider than they are deep. Our finding is in sharp contrast to the narrow and deep trees exhibited by Internet chain letters, as found by Liben-Nowell and Kleinberg [31]. Internet chain letters showed a narrow tree shape (width=82) that went several hundred levels deep (height=288) for a large cascade that involved 18,119 spreaders.

A possible explanation for this discrepancy might be the difference in the way the two systems work. Twitter does not allow its users to restrict the recipients of tweets; tweets are broadcast to all a user's followers. On the other hand, emails can be forwarded to a selective set of users, restricting the propagation to only a fraction of the friends, which creates narrower and deeper cascades.

Related to the temporal dynamics of word-of-mouth, we observed that certain URLs spread for several months (or even several years) in Twitter. This finding could indicate that social media like Twitter and Flickr encourage the interaction of Internet users with media content by forming communities of interest in the World Wide Web. The social media act as channels for distributing content, where users can generate content themselves, discuss it, and rediscover old content. Compared to Flickr or blogs, Twitter showed a much shorter time span for content propagation. The mean time for information to cross each social link was much shorter: 53% of the retweets occurred within a day.

This observation on the long delay is unexpected from the topological structure of the network. Most social networks exhibit the "small-world" property [33], where the average path length between people is so small that every pair of users can be connected in a few hops even when the size of the network reaches planetary-scale. Hence, in theory, this structural property allows new information to spread quickly and widely in the network. The question of which underlying mechanisms lead to such a long delay remains to be answered. We expect to gain insight about such mechanisms from the studies on the bursty dynamics of human behavior [11] and the limited attention span of users [17].

7. RELATED WORK

Throughout the paper, we have discussed the references that closely relate to our work. As our work covers a broad spectrum of topics from information flow models to viral marketing, in this section, we briefly summarize the related work on these topics.

A rich set of theoretical work explains the interplay between the social network structure and information flow. Granovetter [22] proposed a linear threshold model, where someone will adopt an innovation only if a large enough proportion of his neighbors have previously adopted the same innovation. Dodds and Watts [16] studied this model in the field of disease spreading in an epidemiological setting. Watts [43] proposed a mathematical model of global cascades based on sparse Erdős-Rényi random networks and found that global-scale cascade could occur even with few early adopters. Watts examined the conditions for when such cascade happens under homogenous thresholds of user susceptibility. Karsai *et al.* [26] followed the time evolution of information propagation in small-word networks and showed that the slowing down of spreading is found to be caused mainly by weight-topology correlations and the bursty activity patterns of individuals. Steeg *et al.* [37] analyzed information cascades on Digg and concluded that the highly clustered structure of the Digg network limits the final size of cascades observed, as most people who are aware of a story have been exposed to it via multiple friends.

With the advent of OSN data, a number of researchers have presented data-driven analysis and measured patterns of information spreading across social network links. Gruhl *et al.* [24] studied the diffusion of information in the blogosphere based on the use of keywords in blog posts. They presented a pattern of information propagation within blogs using the theory of infectious diseases to model the flow.

Adar and Adamic [8] further extended the idea of applying epidemiological models to describe the information flow and relied on the explicit use of URL links between blogs to track the flow of information. Bakshy *et al.* [10] studied content propagation in the context of the social network existent in Second Life, a multi-player virtual game. By examining cascade trees they find that the social network plays a significant role in the adoption of content.

More recently, Leskovec *et al.* [30] developed a framework for tracking short, distinctive phrases that travel relatively intact through online media. They observed a typical lag of around 2.5 hours between the peak of attention to a phrase in the news media compared to blogs. Similarly, Sun *et al.* [38] also found long chains by studying cascades on Facebook pages, but also showed that these diffusion chains on Facebook are typically started by a substantial number of users. In contrast to these works, our study unveils different aspects of word-of-mouth information, such as not only of the shape of cascades, but also the impact of publishers and subscribers of content. Gomez-Rodriguez *et al.* [20] investigated the problem of tracing paths of diffusion and influence and proposed an algorithm to decide a near-optimal set of directed edges that will maximize influence propagation. Ghosh and Lerman [19] compared a number of influence metrics over Digg.com data and suggested that a centrality-based measure is the best predictor of influence. Scellato *et al.* [34] studied how geographic information extracted from social cascades can be exploited to improve caching of multimedia files in a Content Delivery Network. Their evaluation showed that cache hits can be improved with respect to cache policies without geographic and social information. Wang *et al.* [42] found that social and organizational context significantly impacts to whom and how fast people forward information.

Finally, several recent papers focus on characteristics of the Twitter topology, user influence, and spam. Kwak *et al.* [27] studied the Twitter topology, finding a non-power-law follower distribution, a short effective diameter, and low reciprocity. They also studied approaches for ranking influential users based on the Twitter social network structure. Cha *et al.* [14] showed that highly influential users are not necessarily the most followed users, meaning that aspects of the Twitter topology are not sufficient to capture a user's influence. Galuba *et al.* [18] propose a propagation model that predicts which users are likely to mention which URLs in Twitter. Recent efforts identified different forms of spam [12, 23] and phishing [15] disseminated on Twitter and obfuscated by URL shortening services.

Compared to this body of work, our interest is in investigating the fundamental properties of word-of-mouth-based Web discovery by focusing on URL propagation in OSNs.

8. CONCLUSION

We have presented a first-of-a-kind analysis of word-of-mouth Web discovery using large data gathered from Twitter. Acquisition of such a rich dataset enabled us to reconstruct the cascade processes of various Web URLs and study the role that world-of-mouth based propagation played in making those web links popular within Twitter. Our analysis highlights several important aspects of word-of-mouth based URL discovery, including its impact on URL popularity, dependence on users with a large number of followers, and effect on the diversity of information discovered by users.

There are various directions in which our work can evolve. First, we would like to leverage our findings to improve ranking of actual search engines for real-time content based on the spread patterns of word-of-mouth propagation of URLs. Second, we would like to be able to determine the topological structure of initiators that will speed up the propagation of content in a social network. Such tools would have an important impact in the commercial world, such as advertising and political campaigns, as well as for Web users in general.

9. ACKNOWLEDGEMENTS

Tiago Rodrigues acknowledges the MPI-SWS people for all their support during his summer internship, when this work was done. Meeyoung Cha was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (Grant no. 2011-0012988). We also would like to thank our shepherd Vyas Sekar and our anonymous reviewers.

10. REFERENCES

[1] Alexa. *http://www.alexa.com*.
[2] Bit.ly. *http://www.bit.ly*.
[3] Open directory project. *http://www.dmoz.org*.
[4] Ow.ly. *http://www.ow.ly*.
[5] Rule of three (mathematics) - wikipedia. *http://en.wikipedia.org/wiki/Rule_of_three_(mathematics)*.
[6] tinyurl. *http://www.tinyurl.com*.
[7] Why the open directory (dmoz) is not so open? *http://www.dmozsucks.org/why-the-open-directory-is-not-so-open.php*.
[8] E. Adar and L. A. Adamic. Tracking Information Epidemics in Blogspace. In *ACM Int'l Conference on Web Intelligence*, 2005.
[9] D. Antoniades, I. Polakis, G. Kontaxis, E. Athanasopoulos, S. Ionnadis, E. P. Markatos, and T. Karagiannis. we.b: The web of short URLs. In *Int'l World Wide Web Conference (WWW)*, 2011.
[10] E. Bakshy, B. Karrer, and L. A. Adamic. Social Influence and the Diffusion of User-Created Content. In *ACM Conference on Electronic Commerce (EC)*, 2009.
[11] A.-L. Bárabási. The Origin of Bursts and Heavy Tails in Humans Dynamics. *Nature*, 435:207, 2005.
[12] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting spammers on twitter. In *Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*, 2010.
[13] D. Boyd, S. Golder, and G. Lotan. Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. In *Annual Hawaii International Conference on System Sciences (HICSS)*, 2010.
[14] M. Cha, H. Haddadi, F. Benevenuto, and K. Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy. In *Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*.
[15] S. Chhabra, A. Aggarwal, F. Benevenuto, and P. Kumaraguru. Phi.sh/$ocial: The phishing landscape through short urls. In *Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*, 2011.

[16] P. Dodds and D. Watts. A Generalized Model of Social and Biological Contagion. *Journal of Theoretical Biology*, 2005.

[17] R. I. M. Dunbar. Coevolution of Neocortical Size, Group Size and Language in Humans. *Behavioral and Brain Sciences*, 16(4):681?35, 1993.

[18] W. Galuba, K. Aberer, D. Chakraborty, Z. Despotovic, and W. Kellerer. Outtweeting the twitterers - predicting information cascades in microblogs. In *USENIX Conference on Online Social Networks (WOSN)*, 2010.

[19] R. Ghosh and K. Lerman. Predicting Influential Users in Online Social Networks. In *AAAI Int'l Conference on Weblogs and Social Media (ICWSM)*, 2010.

[20] M. Gomez-Rodriguez, J. Leskovec, and A. Krause. Inferring Networks of Diffusion and Influence. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2010.

[21] Google Geocoding API. `http://code.google.com/intl/en/apis/maps/documentation/geocoding/`.

[22] M. Granovetter. Threshold Models of Collective Behavior. *American Journal of Sociology*, 83(6):1420–1443, 1978.

[23] C. Grier, K. Thomas, V. Paxson, and M. Zhang. @spam: The underground on 140 characters or less. In *ACM conference on Computer and communications security (CCS)*, 2010.

[24] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information Diffusion Through Blogspace. In *Int'l World Wide Web Conference (WWW)*, 2004.

[25] R. Hanneman and M. Riddle. *Introduction to Social Network Methods*. Digital form at http://faculty.ucr.edu/~hanneman/, 2005.

[26] M. Karsai, M. Kivelä, R. K. Pan, K. Kaski, J. Kertész, A.-L. Barabási, and J. Saramäki. Small but slow world: How network topology and burstiness slow down spreading. *Phys. Rev. E*, 83(2):025102, Feb 2011.

[27] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a Social Network or a News Media? In *Int'l World Wide Web Conference (WWW)*, 2010.

[28] K. Lerman and A. Galstyan. Analysis of social voting patterns on digg. In *ACM SIGCOMM Workshop on Social Networks (WOSN)*, 2008.

[29] J. Leskovec, L. A. Adamic, and B. A. Huberman. The Dynamics of Viral Marketing. *ACM Trans. on the Web (TWEB)*, 2007.

[30] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-Tracking and the Dynamics of the News Cycle. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2009.

[31] D. Liben-Nowell and J. Kleinberg. Tracing Information Flow on a Global Scale using Internet Chain-Letter Data. *Proc. Natl. Acad. Sci. USA*, 2008.

[32] D. Liben-Nowell, J. Novak, R. Kumar, and P. R. A. Tomkins. Geographic Routing in Social Networks. *Proc. of the Nat. Academy of Sciences (PNAS)*, 102:11623–1162, 2005.

[33] S. Milgram. The Small World Problem. *Psychology Today*, 1(3), 1967.

[34] S. Scellato, C. Mascolo, M. Musolesi, and J. Crowcroft. Track Globally, Deliver Locally: Improving Content Delivery Networks by Tracking Geographic Social Cascades . In *Int'l World Wide Web Conference (WWW)*, 2011.

[35] E. Schonfeld, Facebook Drives 44 Percent of Social Sharing On The Web, *TechCrunch*, 2010. *http://techcrunch.com/2010/02/16/facebook-44-percent-social-sharing/*.

[36] O. Sorenson. Social networks and industrial geography. *Journal of Evolutionary Economics*, 13(5):513–527, 2003.

[37] G. V. Steeg, R. Ghosh, and K. Lerman. What Stops Social Epidemics? In *AAAI Int'l Conference on Weblogs and Social Media (ICWSM)*, 2011.

[38] E. Sun, I. Rosenn, C. Marlow, and T. Lento. Gesundheit! Modeling Contagion Through Facebook News Feed. In *AAAI Int'l Conference on Weblogs and Social Media (ICWSM)*, 2009.

[39] G. Topa. Social interactions, local spillovers and unemployment. *Review of Economic Studies*, 68(2):261–95, 2001.

[40] A. Campbell. Social Activity Becomes Significant Source of Website Traffic, *Small Business Trends*, 2009. *http://smallbiztrends.com/2009/03/social-activity-significant-source-website-traffic.html*.

[41] L. Rao, Twitter Seeing 90 Million Tweets Per Day, 25 Percent Contain Links, *TechCrunch*, 2010. *http://techcrunch.com/2010/09/14/twitter-seeing-90-million-tweets-per-day/*.

[42] D. Wang, Z. Wen, H. Tong, C.-Y. Lin, C. Song, and A.-L. Barabasi. Information Spreading in Context. In *Int'l World Wide Web Conference (WWW)*, 2012.

[43] D. J. Watts. A Simple Model of Global Cascades on Random Networks. *Proc. of the Nat. Academy of Sciences (PNAS)*, 2002.

[44] Z. Yang, J. Guo, K. Cai, J. Tang, J. Li, L. Zhang, and Z. S. Understanding Retweeting Behaviors in Social Networks. In *ACM Conference on Information and Knowledge Management (CIKM)*, 2010.

Summary Review Documentation for

"On Word-of-Mouth Based Discovery of the Web"

Authors: T. Rodrigues, F. Benevenuto, M. Cha, K. Gummadi, V. Almeida

Reviewer #1

Strengths: Timely, interesting problem. Great dataset. Well organized. Interesting insights.

Weaknesses: I was less convinced about the tie-in to networking/systems. Also, its hard to judge the novelty given that most of the work in this area happens outside core IMC/networking venues.

Comments to Authors: This paper was a lively interesting read, with well motivated questions. I do not see any major problems, but have some minor comments/issues:

1. My understanding was that the Twitter API only allows a "random" sample approximately 5% of the tweets. Is your data sampled in any way; if yes how does that bias the findings? Or did you obtain a more exclusive access; if so please clarify.

2. The fact that implicit links uncover 24X more links raises a few red flags; can you provide some more analysis/insights into why this is the case or at least validate this better?

3. "58 whitelisted servers" – I do not follow what you mean here; maybe this is related to (1)

4. The exponents for indegree/outdegree seem to be very similar; this is surprising since all anecdotal evidence suggests that the twitter graph is highly asymmetric (maybe there is no discrepancy here because the power law exponent does not tell anything structurally, but I was intrigued).

5. Early on in the paper, you refer to "trees", when in fact the propagation structure is a forest; maybe you want to be more careful.

6. Again, I found that 80% of people retweet the most recent post weird; is this an effect of specific user interface effects or human behavior in "batching" access to reading tweets?

7. Maybe this is obvious (e.g., look at timestamps?), but it would help to spend a small section describing your heuristic for identifying "implicit" links. Also, I assume that the 80% is for explicit retweets -- does that extrapolation also apply to the implicit links also?

8. I do not understand why you specifically look at the largest tree in Sec. 4.3; why not look at all propagation trees instead of just the largest component in the forest?

9. I had a tough time understanding the discussion in the first 2 paragraphs on page 11. At one point you say most retweets occur on day 1, but then the next paragraph talks about long delay. What timescales are you talking about here?

Reviewer #2

Strengths: Interesting observations are presented and the paper is well written. The difference in the popular domains on Twitter and on the web is presented. The result verifies that Twitter create chances for niche URLs to become popular. The results on content locality of the spreads are new.

Weaknesses: Some of issues have been studied in the paper of WSDM2011 by Bakshy et al., e.g., type of URL vs. cascade size, content category vs. cascade size, height of the cascades.

Comments to Authors: There are overlaps between this paper and the WSDM paper mentioned above. This paper also includes new results, which are interesting considering the location aspects. It is interesting to see that the audience size of less popular domains can be comparable to those of the popular ones, which suggests that niche URLs can be popular on Twitter.

The claim that as twitter becomes popular source of information, niche website can be popular. However, it is not clear that despite the popularity of twitter, web search might still be important and dominate source of web content discovery.

Some more comments:

- Regarding Section 5.2 the expected traffic pattern, why videos traffic are considered in hours and not the exact size of the videos? Also, the suggestion about the local server for content distribution may need further investigation as Twitter is not the only source of traffic for websites.

- Section 6 (1), the third paragraph discussing the relationship between the in-degree distribution and cascade size distribution is hard to follow. What does "collaborative act of sharing" mean? How does it relate to the two distributions?

- Section 6.1, in the last paragraph discussing the redundant information, the authors say that for URLs with > 100 spreaders, at least one users receive the URL multiple time. This does not seem significant in comparison with the case when the URLs with fewer spreaders. Is there stronger support?

Reviewer #3

Strengths: Information propagation has become an important area in OSNs and hence the paper is interesting in that respect. The authors argue that Twitter propagation trees may be different than what previous work has shown for more traditional media such as email. The paper is well written.

Weaknesses: The paper appears more as a preliminary study of URL forwarding without going deep into the topic. No justifiable explanations or models are explored for the particular observations (i.e., width of propagation trees), while the data is there. Some parts of the paper are known observations from previous work.

Comments to Authors: This is a well-written, easy to follow paper on an interesting topic - content propagation in Twitter. The main point of reference is forwarding of URLs which indeed is a major component in Twitter.

The paper does present a couple of novel observations, with the most interesting of those being the fact that twitter propagation trees are more wide than deep. Similarly, the analysis on content distribution based on location is interesting - although projecting this to traffic patterns does not appear as convincing. In my view, from the 5 key findings listed in the introduction, the 4th and 5th are the most interesting (and novel) ones.

However, it is disappointing that the paper only remains superficially at these observations without delving deeper regarding the factors that lead to these observations. For example, the authors should have examined what contributes most to this effect (i.e., large width), and for which types of users - instead only conjectures are made referring to the follower-mechanism of twitter as the main cause.

Besides this, I have a couple of other reservations with the paper:

First, the authors concentrate a lot on the differences observed in the height and width of the propagation trees, when compared to the chain-letter work by Nowell and Kleinberg [30] (key finding number 4 in the introduction). While indeed the differences appear true, the methodologies generating the corresponding propagation trees in each case are different. It is thus unclear how much this comparison makes sense. The authors assume that each node's parent is the last node sending the content, while in [30] a spanning tree algorithm is run. Does this result to any differences in the shape of the trees? Could the authors use the same tree generation algorithm as in [30] and then compare the observations?

Second, the authors present a number of observations that have been presented in previous work, without however the authors sufficiently comparing or referring to it. Two examples here include: 1) The fact that distinct initiators result in large cascades (section 4.1) has been observed before in Digg (WOSN 2008, Lerman et al. - not cited in the paper). 2) URL propagation in twitter was also studied in the context of short URLs in [9] with newer datasets. [9] observes exactly the third key finding of the authors mentioned in the introduction. In fact, the analysis in most of Section 3 is the same as in [9] (without this being mentioned), but the findings are quite different in some cases (e.g., percentage of bit.ly URLs) which might reflect the change between 2009 (traces used in the paper) and 2010 (data in [9]). It would be interesting to cross-reference the results of the two papers, which is avoided in the paper. Yet, the fact remains that these are not as novel observations as the ones regarding the propagation trees.

Other comments:

The discussion section is quite lengthy and it is unclear that it offers much information for the rest of the paper.

Section 2.1: It would be nice to provide definitions for the in-degree and out-degree.
Section 2.2: What do you mean by URL recycling, and where is the "120 days of inactivity" coming from? Reference?

Section 3.4: Is the 17% sample enough for this categorization? DMOZ is often criticized for its representativeness.

Section 4.1: It would be nice to provide also the p-value for the correlations.

Section 4.3: It would be interesting to see the results with other definitions of width besides the maximum.

Section 4.4: The example in the last paragraph is not convincing to me, since if the email is forwarded to all A-E, the resulting tree the same as in twitter. People do have e-mail lists where they forward fun and other stuff (e.g., images, Youtube links etc), where the resulting propagation trees might be similar to what you observe for Twitter. This goes back to the main comment on the paper - the origins of the width are not sufficiently examined.

Section 5.2: How do you know that users indeed looked at the content and not blindly re-tweeted?

Reviewer #4

Strengths: - Large data set that is studied from many angles.
- Study of impact of geolocation on propagation.

Weaknesses: - No study of the propagation time.
- Data set is almost 2 years old.

Comments to Authors: This is a clear paper that is nice to read. I think you did a good job at looking at the data set from different angles to characterize word of mouth; the main dimension that is absent is a study of the propagation time of these words of mouth. But the results are really incremental.

It is expected that when you have a broadcast mechanism available like Twitter, you will create a tree that is wider than if you forward a chain letter and have to choose who you want to forward the email to.

What would make the paper more exciting is to compare, if possible, the word of mouths of different OSNs: e.g. Twitter vs. Facebook.

I like your comparison of popular content with Alexa and your study of word of mouth for different content type.

How would Figure 5 look like if the ranking of domains would also be based on the overall audience instead of the number of links. I am not sure that I am convinced that it makes sense to have a ranking based on the number of URLs. You should first study the correlation between overall audience and number of links.

"Considering that content in Twitter is typically produced by geographically-diverse users but consumed locally, one could allow users to upload content to a local server in the corresponding geographical area, like a city. Such a mechanism could significantly reduce the amount of wide-area bandwidth needed compared to an upload to a centralized, remote server. Additionally, the trend towards local consumption and the fact that users are uniformly distributed across different locations indicates that placing a server in every region can reduce the amount of cross-region traffic." - You are simply saying that CDNs would make the distribution more efficient which is not really surprising, and companies like YouTube/Google are already doing that.

Reviewer #5

Strengths: The topic is interesting, as it works toward to goal of understanding what makes content propagate quickly/widely in a social network. The paper confirms some widely held beliefs, and is well written and well executed.

Weaknesses: I quite dislike two aspects of this paper. First, it is too long. I would have much preferred the findings in a crisp 7-pager than the given 14 pages. Second, the paper is typical for IMC in that a lot of the findings are not surprising at all yet the authors tend to make it sound like they are.

Comments to Authors: "Can word of mouth reach a wide audience?" Obviously. It is called viral marketing. I find it similarly obvious that widely spread URLs can originate on unpopular websites - it matters who spreads the content and how, not where it comes from. That too is central to viral marketing. It is scientifically completely valid to confirm the intuition, but there is no reason to sound surprised (e.g. "it is striking", "interestingly") when you do.

The analysis of Table 4 is brief. You say it covers a diverse set of topics. I would argue that it reflects rather clearly that you're looking at Twitter, given that around half of the top 10 are clearly related to it. What is #2, the Facebook warning page? (Change the caption to "Top 10 URL domains [...]".)

In subfigure 6a, how can the rank of a URL be 1/10, i.e., less than 1? How many initiators did the most popular URL from that plot have?

How often is the initiator the contributor of the widest point in a cascade? I.e., does Lady Gaga find new stuff, or merely retweet?

The comparison to chain letters feels tedious. Email and Twitter are substantially different communication media; it is completely clear that chain letters are more narrow yet deeper. As you are saying, Twitter is inherently broadcast, but chain letters typically explicitly ask for forwards to a given (small!) number of people.

I wanted the paper to end on page 9 -- everything except for the related work section after it feels like it is inflating the paper to full length. I really do not see how Sec 6.1 is part of "Discussion". I think a terrific way to make this a better full paper would have been to integrate the analysis of the social structure of the graph into the URL propagation patterns. That would shed some light into *why* content propagates, not merely what content, and how.

Response from the Authors

We thank all the reviewers for their great suggestions. We tried to incorporate these suggestions in the camera ready version whenever possible. In particular, we fixed small mistakes and clarified description of the analyses throughout the paper. We cited the related works suggested, comparing with our analysis whenever possible (e.g., in sections 3.2 and 4.2). We also reduced the discussion section as one of the reviewers suggested. We believe these changes increased the quality of our paper.

In some parts of the suggestions the reviewers did not agree among themselves, in which case we decided not to apply any change. For example, reviewer #3 suggested that we extend the analysis on the comparison of cascade shapes with that of the Internet chain-letters, while reviewers #4 and #5 commented that the differences are obvious and expected.

We also left some of the suggestions (e.g., providing new models of cascades, adding new deeper analysis on cascade shapes) as future work, as they are interesting but there was not time to make these changes. We plan to address these suggestions when we extend our work for a journal submission.

In conclusion, we appreciate all the important suggestions the reviewers provided in helping us prepare a high quality ACM IMC paper.