Characterizing Podcast Services: Publishing, Usage, and Dissemination

Dinan Gunawardena, Thomas Karagiannis, Alexandre Proutiere and Milan Vojnovic
Microsoft Research
Cambridge, UK
{dinang, thomkar, aproutie, milanv}@microsoft.com

ABSTRACT

In this paper, we aim at characterizing podcast services both from publishers' and users' perspectives, and at analyzing the implications of these characteristics on the design of efficient dissemination systems. Specifically, our goal is to characterize how podcasting content is generated and published, and how users subscribe and consume podcasts. We are also interested in understanding whether podcast episodes are efficiently disseminated to users just using a sporadic direct access to the Internet (which is the current way of downloading podcast episodes), or whether the use of peer-to-peer mobile device-to-device dissemination systems could help enhancing the performance of podcast services.

Our study is based on traces of podcast episode releases, subscriptions, and play times from major podcast service providers. An extensive analysis of the traces allows us to develop a comprehensive model of current podcast services, and provide statistics about the type and content of the typical podcasts, the size and the release frequencies of their episodes, as well as their popularity. By studying podcast usage, we show that the service is delay-tolerant, as users may well play podcast episodes a long time after their actual release. An interesting consequence of this delay tolerance is that mobile device-to-device dissemination systems would not be very useful for the current typical podcasts, while they may become more attractive for future interactive podcast services.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: general; C.4 [**Performance of Systems**]: design studies

General Terms

Human Factors, Measurement

IMC'09, November 4–6, 2009, Chicago, Illinois, USA.
Copyright 2009 ACM 978-1-60558-770-7/09/11 ...$10.00.

1. INTRODUCTION

Podcasting has evolved into one of the mainstream Internet services. Podcasts refer to series of files (i.e., feeds), typically of media type such as audio or video, that are distributed through web syndication mechanisms such as RSS (Really Simple Syndication). Users subscribing to a podcast retrieve the corresponding files or *episodes* when they are made available by accessing a centrally-maintained web feed. The episodes are thus automatically downloaded and then stored locally on the user's computer or other mobile device, for offline use. Recent studies suggest that the podcast audience comprises at least one in ten Internet users today while it is expected to double over the next five years [1]. Traditional popular media and content providers expand their audiences through podcasts, and online marketers use podcasting to increase sales as correlation between podcast listeners and online shopping has been observed [2]. However, despite this widespread growth, studies of the characteristics of podcasting services have been limited, see for example [3].

This work is a step towards advancing our understanding of podcasting services along three dimensions: Publishing, usage and dissemination. In particular, our analysis provides answers to the following questions.

Q1. Publishing: Who are the main podcast creators? Where are podcast episodes typically published? What are the typical content and size of podcast episodes? What are the underlying dynamics of podcast content generation?

Q2. Usage: How do users subscribe to podcasts and consume content? What are the most popular podcasts? Can we identify potential models explaining how users subscribe to the different podcasts and leading to the observed discrepancy in their popularity?

Q3. Dissemination: What is the typical delay between the times podcast episodes are released and played by users on their PC or mobile device? Given this delay and the statistical characteristics of podcasts, what are the most natural and efficient ways to disseminate the podcast episodes to users? When could ad-hoc mobile device-to-device dissemination be of interest?

Our analysis is based on a 70-day dataset, comprising logs from one of the major podcast provider services. In contrast to previous works that focused on properties of content generation through crawled data, our dataset further captures user actions such as subscription, unsubscription and podcast consumption (i.e., play) events. This is important as it allows us to distinguish "true" user interest by examining which of the user subscribed podcasts are actually consumed. Additionally, our data provides visibility into the evolution of podcast popularity both from a user subscription (i.e., *What's popular?*), and from a user play perspective (i.e., *What's hot?*). To stress-test the representativeness of our data, we also crawl additional podcast provider services.

Analyzing the podcast services from the publishers' perspective, we find that podcasters may either be traditional broadcasters such as TV and radio channels, but may also be independent creators publishing their podcasts via one of the main podcast portals, e.g., Google's feedburner. We also observe that podcast content is generated periodically, with the most prominent periods being the weekly and daily content generation. The median episode size is roughly 15 Mbytes for audio and 30 Mbytes for video, and the median rate of content released per podcast is about 20 Mbytes per week, but both episode size and rate depend on the type of podcast (i.e., audio or video). We further find that content does not appear to be released uniformly throughout the day.

Regarding podcast usage, we find that while the popularity of podcasts across different perspectives (podcast vs. play) follows in general the Pareto principle (i.e., 80-20 rule), only roughly 20% of the rankings of podcasts based on subscriptions and plays coincide. This effect appears invariant across the set of most popular feeds, and holds for roughly the top-100 most popular podcasts. This finding implies that subscription popularity does not necessarily imply consumption of the podcast content. In general, we observe that users subscribe to roughly 6 podcasts on average, while at the same time consume less than 4 per week. Subscriptions per podcast on the other hand appear to grow linearly in the number of existing subscriptions. Our observations suggest existence of "rich-get-richer" type of popularity reinforcements, which might be the effect of selective podcast promotion through the user interface (i.e., display of a "what's popular" or "what's hot" list). Content consumption exhibits a dichotomy with respect to the device used, with users playing podcasts typically on a single type of device, a PC or a mobile device, but not on both. This observation suggests that users develop a consistent way of interacting with the service.

Finally, concerning podcast dissemination, one of our main findings is that podcasts currently constitute a delay-tolerant service. Specifically, the median delay between the times episodes are released and played by users is about 10 days and that this delay is 1 day or shorter for only as few as 1% of podcasts. Given this delay-tolerance and the statistical characteristics of podcasts, it seems that for most of the podcasts, it would suffice to synchronize podcast contents on mobile devices a few times a day. Thus, given the current nature of the podcasts, their dissemination may be done through sporadic access to the Internet (e.g., while the device is connected to a PC when the user is at home or office). Mobile device-to-device dissemination would be of interest for podcasts whose delay tolerance is less than a day (as otherwise, the content can be synchronized while connected to the Internet). While we didn't find such podcasts to be typical, they may well become popular in the future. This is indeed suggested by the rising trend towards more interactive podcasting-like services, e.g., twitter. Device-to-device dissemination could also be of interest in scenarios where the Internet access is limited, e.g., while on travel or at places with limited infrastructure.

Our contributions can be summarized in the following points:

- We provide an extensive characterization of a popular podcast provider service based on a 70-day dataset. This analysis allows us to suggest generic models for the creation and the publication of podcasts, as well as for the way episodes belonging to the same podcast are released (Section 3).

- We describe possible models for user-to-content matching by examining how users may subscribe to podcasts and play their episodes. The analysis reveals that popularity reinforcements are in place (Section 4).

- We provide an initial discussion on the implications of our findings for podcast dissemination systems. Specifically, it seems that disseminating podcasts through classical sporadic access to the Internet suffices given their current delay-tolerant nature. Ad-hoc mobile device-to-device dissemination could become more appropriate for future interactive podcasting services (Section 5).

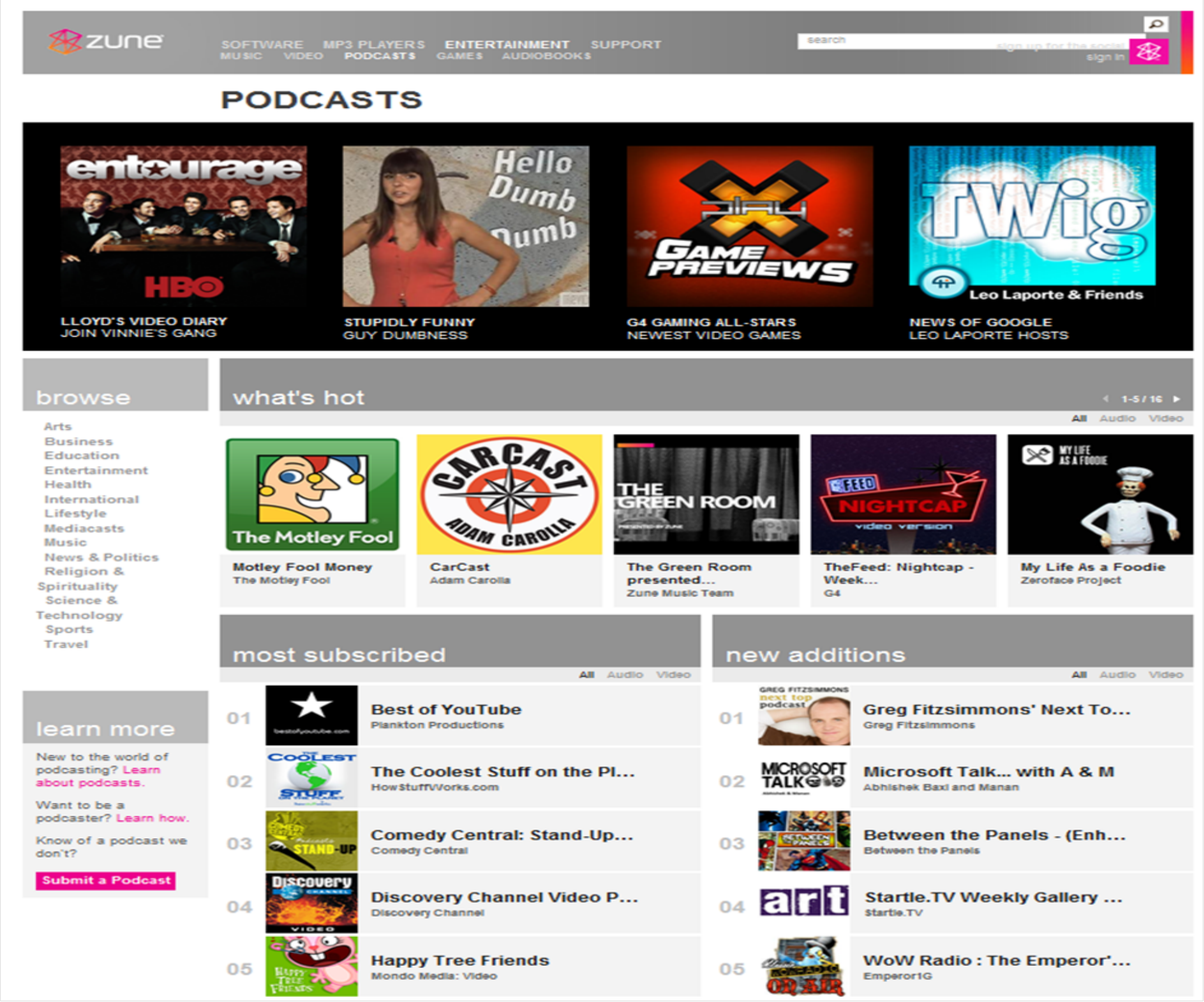
We believe that our study is among the first to provide guidelines for efficient design of large scale podcast services by examining several properties, both from a service and also from a user perspective. Our findings are further informative for the design of dissemination systems, such as ad hoc podcast services [4], and in general, for modeling and simulations in performance evaluation studies in the context of publish-subscribe systems.

2. DATA SETS AND METHODOLOGY

The results presented in this work are mainly based on a 70-day trace from the Zune Social service [5]. The service provides a web interface for users to discover, download, play and purchase media items, such as songs or videos, games, audiobooks and podcasts. The service also offers social networking features, such as exposing playlists to friends. Media items can be played either on regular PCs through the Zune software, or could be uploaded to and played on the Zune device.

Our focus is on podcasts. Fig. 1 presents a snapshot of the Zune Social user interface for the podcasting service. Users are presented with featured podcasts in three main lists (i.e., *what's hot?*, *most subscribed*, *new additions*), and additionally they may browse available podcasts classified under a number of categories or through a search service. Users may subscribe or unsubscribe to podcasts, and are allowed to sync and play podcasts to which they have subscribed. Our data captures all such actions. Specifically, each log entry specifies a timestamp, a user and a media id (both ids are hashes), and the corresponding action (subscribe, unsubscribe and play). For play actions, our data further indicates whether plays occur on the PC client or on the Zune device. Overall, in our dataset, we observe a total number of podcast users in the order of several hundred thousands and more than 8 thousands podcasts.

To infer the media type (e.g., audio or video) as well as acquire information regarding the files released per podcast (henceforth referred to also as episodes), we perform a crawl



PODCASTS

LLOYD'S VIDEO DIARY
JOIN VINNIE'S GANG

STUPIDLY FUNNY
GUY DUMBNESS

G4 GAMING ALL-STARS
NEWEST VIDEO GAMES

NEWS OF GOOGLE
LEO LAPORTE HOSTS

browse

Arts
Business
Education
Entertainment
Health
International
Lifestyle
Mediacasts
Music
News & Politics
Religion & Spirituality
Science & Technology
Sports
Travel

learn more

New to the world of podcasting? Learn about podcasts.

Want to be a podcaster? Learn how.

Know of a podcast we don't?

Submit a Podcast

what's hot

Motley Fool Money
The Motley Fool

CarCast
Adam Carolla

The Green Room presented...
Zune Music Team

TheFeed: Nightcap - Week...
G4

My Life As a Foodie
Zeroface Project

most subscribed

01	Best of YouTube Plankton Productions
02	The Coolest Stuff on the Pl... HowStuffWorks.com
03	Comedy Central: Stand-Up... Comedy Central
04	Discovery Channel Video P... Discovery Channel
05	Happy Tree Friends Mondo Media: Video

new additions

01	Greg Fitzsimmons' Next To... Greg Fitzsimmons
02	Microsoft Talk... with A & M Abhishek Baxi and Manan
03	Between the Panels - (Enh... Between the Panels
04	Startle.TV Weekly Gallery ... Startle.TV
05	WoW Radio : The Emperor'... Emperor1G

Figure 1: A snapshot of the Zune Social podcasting service. Users can subscribe to podcasts by choosing from three pre-compiled lists, namely, *"what's hot?"*, *"most subscribed"*, and *"new additions"*. Additionally, users can browse podcasts by category or search for podcasts by keywords.

of the Zune Social service. Crawling provides specific podcast information from the RSS xml file, such as the mime type, title and generation time of podcast episodes, file size, etc. The media id allows us to correlate trace entries with the information crawled. However, care needs to be taken as we observed several discrepancies in the xml fields, especially regarding the episode size with missing or incorrect values. The size field appears to be manually entered by the podcaster operating the feed. To correct for such discrepancies we further infer the size of podcast episodes by actually downloading the episodes. We observe a significant fraction of incorrectly reported file sizes (about 35%), and for about 13% of the files the relative error is larger than 10%. We further manually categorize podcasts by extracting frequent keywords from the xml files, following the categories provided by the Zune service (see Fig. 1, i.e., Arts, Business, Education, Entertainment, Health, International, Lifestyle, Mediacasts, Music, News & Politics, Religion & Spirituality, Science, Technology, Sports and Travel).

Since our dataset is a 70-day snapshot of the service, we cannot observe events that took place before the tracing period. This creates some inconsistencies in user action sequences, by, for example, observing a play action or an unsubscribe event without observing a corresponding subscription event. To account for these inconsistencies, we identify all such actions, and insert subscription events for all such users and the corresponding media ids at the beginning of the trace (i.e., at time 0). This correction allows us to infer the true subscription popularity per podcast. Note that when examining dynamics such as the subscription rate per podcast in the following sections, we leave all such inserted subscriptions out of the analysis, as they could bias findings (e.g., creating a spike of subscriptions at time 0).

Finally, to examine the representativeness of the Zune data set, we crawl the top-100 podcasts from additional podcast provider services, namely iTunes US and UK. However, note that this comparison is essentially feasible only for characteristics that are visible through the RSS xml files, such as the podcast episode release times, file sizes, etc.

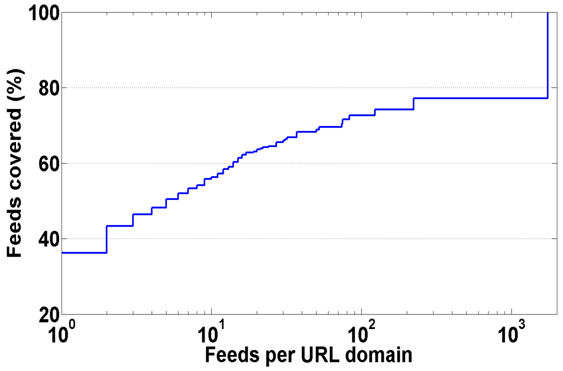


Figure 2: CDF of the number of feeds per url domain (e.g., 55% of the feeds are published on url domains publishing at most 10 feeds).

3. PUBLISHERS AND CONTENT

In this section, we present a statistical analysis of the podcast service from the publishers' perspective. Specifically, we first identify typical podcast publishers. For example, are podcasts published by traditional broadcasters such as hbo, bbc, nyt, or by independent creators? Then, we categorize podcasts according to the mime type (audio, video, application, or other) of the corresponding files and the type of contents (news, music, entertainment, etc.). We finally analyze how episodes of a given podcast are released in time, and characterize the sizes of these episodes depending on the type of the corresponding podcast.

3.1 Publishers

Podcasts are created and published by traditional broadcasters, such as TV/radio channels or newspapers, but also by more independent creators. The big names of broadcasting usually publish their podcasts on their respective websites, see e.g., www.hbo.com/podcasts, whereas independent podcasters may publish their podcasts also on social networking websites (either specialized in podcasting, e.g., feedburner.google.com, mypodcast.com, or not, e.g. facebook.com). Fig. 2 shows the distribution of the number of podcasts published on a single url domain[1] (i.e., representing a single publisher) for the Zune data set. Url domains publishing a single podcast cover 37% of the podcasts, whereas 25% of podcasts are published through a single publisher, namely feedburner.google.com (previously feeds.feedburner.com). The second and third most important url domains in terms of the number of podcasts correspond to radio websites (npr and bbc), but they publish roughly 10 times less podcasts than feedburner. Interestingly, we observe that the url domain does not influence the number of user subscriptions a podcast receives. In other words, the percentage of user subscriptions per podcast appears independent as to where the podcast is published. This implies that publishing a podcast in a high-profile aggregator site does not provide advantages with respect to the audience population. We extensively study subscription properties in Section 4.

Table 1: Distribution of mime types of files.

	Audio	Video	Other	Application
Zune	74.5	15.9	8.7	0.9
iTunes US	60.4	37.8	1.4	0.4
iTunes UK	57.5	34.9	7.1	0.4

Table 2: Distribution of content type per feed.

	Zune		iTunes US	iTunes UK
	all	top 100	top 100	top 100
News	29.2	17.6	11.6	11.2
Entertainment	25.8	29.6	20.9	26.4
Music	17.9	9.3	17.7	16.5
Mediacasts	15.8	6.5	8.4	7.0
Education	12.3	7.4	4.2	2.5
Science	10.7	12.0	4.2	5.0
Religion	10.2	0.9	3.2	2.5
Lifestyle	8.2	7.4	9.3	10.3
Business	6.8	0.9	7.0	4.1
Sports	5.4	2.8	4.2	4.1
Travel	5.0	1.8	2.8	2.9
Technology	3.9	2.8	0.0	0.0
Health	3.6	0.0	1.4	1.2
Arts	2.6	0.0	1.8	2.0
International	0.7	0.9	3.2	4.1

3.2 Content of Podcasts

The very first podcasts were meant to put prerecorded radio programs online. Today, the majority of podcast files are still audio files; however, we observe an increasing diversity both in the type of files released by podcasters, and in the type of contents. Table 1 shows the fraction of files of the different mime types. Video files are the second most podcasted types of files. Note that in the Zune and iTunes UK data sets, the mime type "Other" represents a non-negligible fraction of the set of files - we were unfortunately unable to identify the actual type of most of these files. The mime type "Application" is rare and corresponds, in most cases, to text documents (pdf, ms word) and to binary files (most likely sildeshows). We have also manually categorized files from the three data sets depending on the type of their content. The categorization was based on keywords of the xml files according to the categories provided by the Zune Social service (see Section 2), with less than 30% of podcasts (per data set) not being classified. The results are reported in Table 2. Entertainment, music and news podcasts represent between 50 and 75% of the podcasts depending on the data set considered. Regarding the Zune service, the most common types of podcasts depends on whether all podcasts or the 100 most popular podcasts are considered: overall news is the most represented content, but among the most popular podcasts, files of entertainment content are dominant.

Surprisingly, we have discovered that a single podcast may release files from different mime types. This observation is reported in Fig. 3, where we plot the fraction of the most

[1] A url domain is identified by the first *block* of the url name. For example, "edition.cnn.com/tech" and "edition.cnn.com/services" belong to the same domain, but not to the same domain as "weather.edition.cnn.com/".

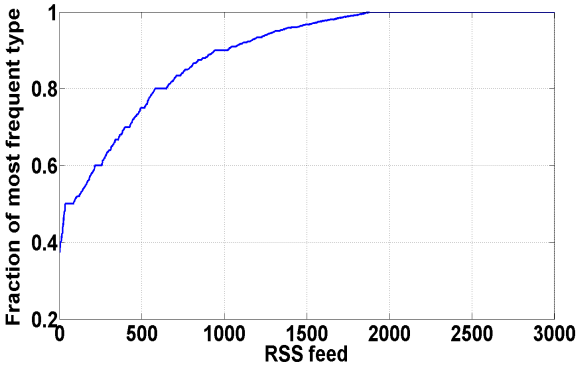


Figure 3: Diversity of mime types of files released by a single podcast. 25% of podcasts publish episodes of different mime type.

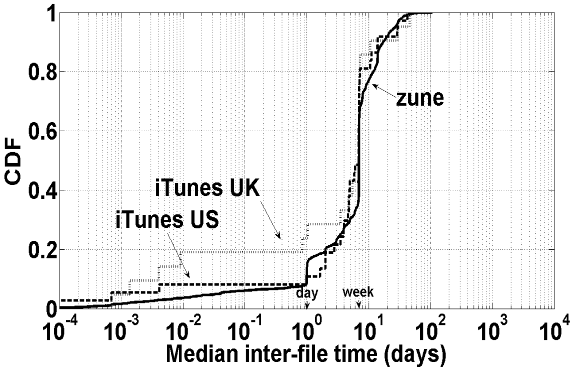


Figure 4: Inter-file release time CDF for the three datasets. 30% of podcasts publish episodes on a weekly basis.

frequent mime type of released episodes per podcast. One fourth of podcasts publish files of different types. There are basically two reasons for this: First, some podcasts do actually release episodes of different types, video and audio; then, there are for some files problems of classification, e.g., files with mp4 extension are classified as video, whereas they might also be audio files.

3.3 Generation of Podcast Episodes

We now analyze how episodes from the same podcast are released over time. In Fig. 4, we provide the Cumulative Distribution Function (CDF) of the median inter-file release time across the various podcasts. For all three data sets, a lot of podcasts (about 30%) generate and release episodes weekly, and about 10% of the podcasts release episodes daily. In Fig. 5, we test whether the frequency of release of episodes is correlated with the size of these episodes. This does not seem to be the case: podcasts generating high-volume episodes. i.e., videos, have statistically similar release frequencies to audio podcasts. We find that episode releases are spread evenly over working days (about 17% of the episodes are released each of these days), and over week-end days (about 7-8% of episodes are published on Saturday or Sunday). In Fig. 6, we show the fraction of episodes released at a given time during the day. Releases are roughly uniformly spread during the day, but exhibit higher activity around midnight (4 to 6 AM UTC) and around noon (3 to 6 PM UTC).

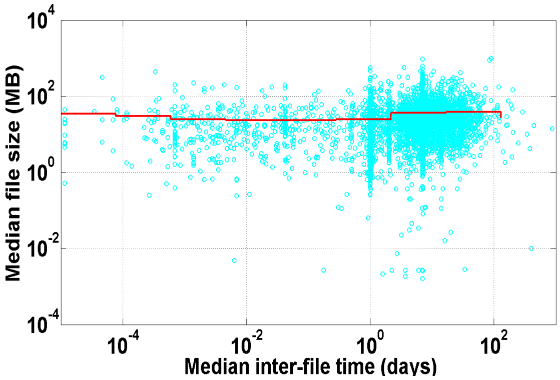


Figure 5: Median file size vs. median inter file-release time. The circles represent per-feed values while the solid line is the mean value of the median file sizes over log-binned median inter-file times. Inter-release time does not appear correlated with the file size.

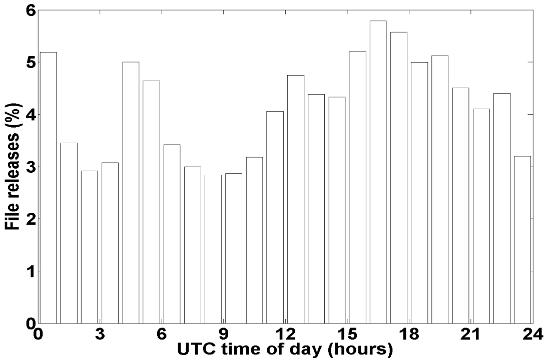


Figure 6: Distribution of the file release time within the day. Higher podcasting activity occurs during 4-6am and 3-6pm UTC.

3.4 File Sizes and Rates of Podcasts

We now examine the statistical properties of the sizes of released episodes, and those of the rates of podcasts. Fig. 7 shows the range of the size of files for the same podcast. We observe that very few files exceed 100 Mbytes - which also happens to be the maximum file size recommended for YouTube videos. Fig. 8 presents the CDF of the sizes of files of given mime type across all files. The median size of video files is about 30 Mbytes, whereas the median size of audio files is rather close to 15 Mbytes. These numbers are consistent with past studies that estimated a median size of 22 MBytes [3]. Computing the file size CDF per podcast (Fig. 9) results in similar observations. Table 3 summarizes the median and mean file size across the various types.

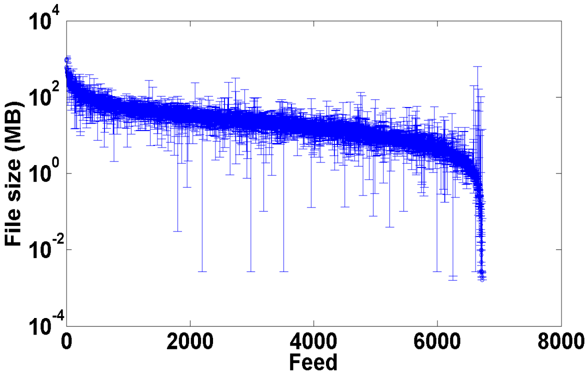


Figure 7: File size range per feed. Most files are less than 100 Mbytes.

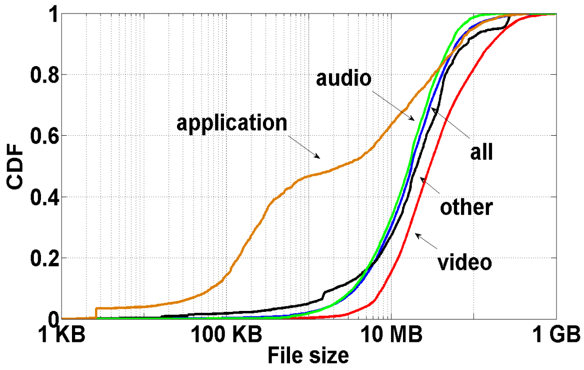


Figure 8: File size CDF across different mime types. The median is around 30Mbytes and 15Mbytes for video and audio files respectively.

We also investigate whether podcasts deliver episodes of the same sizes: Fig. 10 presents the CDF of the coefficient of variations of the files within a given podcast. We find that 80% of the podcasts release always files of similar sizes - a coefficient of variation less than 2. Podcasts publishing files of very different sizes coincide with podcasts delivering files of different mime types.

Finally, we examine the publishing rates of podcasts in terms of Mbytes and files per week in Figures 11 and 12 respectively. The figures show that typically podcasts publish roughly a file per week or around 20 Mbytes worth of data per week. These numbers are especially relevant for dissemination scenarios, as they provide some hints on what the required synchronization frequency of content might be.

4. USAGE

We now focus on podcast usage and how users subscribe to podcasts and play podcast episodes. Specifically, we provide a characterization of the following properties.

- Podcast subscriptions and consumption of podcasts by users over different timescales.

- The existence of podcast popularity reinforcements.

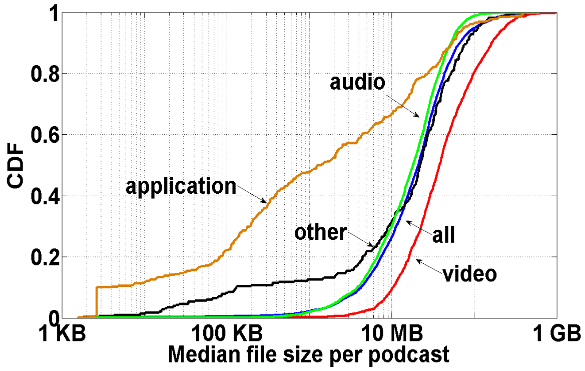


Figure 9: File size per podcast across different mime types. The distributions appear similar to the ones in Fig. 8.

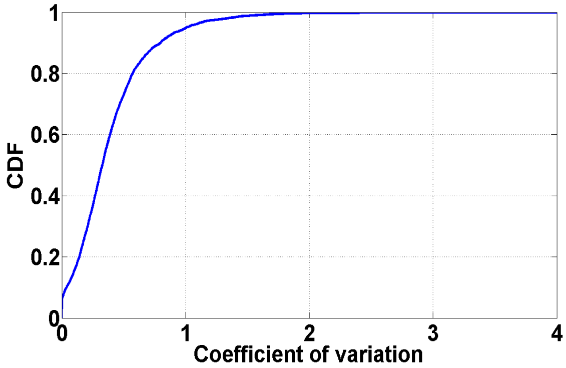


Figure 10: File size coefficient of variation per podcast. 80% of the podcasts release episodes of similar sizes.

- The existence of user information gateways bringing new podcasts into the service.

4.1 User Subscriptions

The service has experienced a significant growth during the period of the study. Fig. 13 presents new users observed as a cumulative fraction of the total population over the whole dataset. The figure highlights that the population of podcast users increased by roughly 250% during the 70 days of the study. We observe an increase in the overall subsscription rate around day 37 which we have attributed to changes in the interface of the Zune Social service. Note that 40% of the users appear at day 0 in the figure, which is a side-effect of inferring existing subscriptions to podcasts (see Sec. 2).

Fig. 14 shows the distribution of the number of subscriptions per user which has a median value of 3 and it exceeds 40 for only as few as 1% of users. The mean number of subscriptions per user is double the median indicating a skewed distribution where a few users have many subscriptions and most have a few. The number of subscriptions per user is especially relevant for dissemination scenarios as it reflects the requirements on transfer rates and storage to synchronize

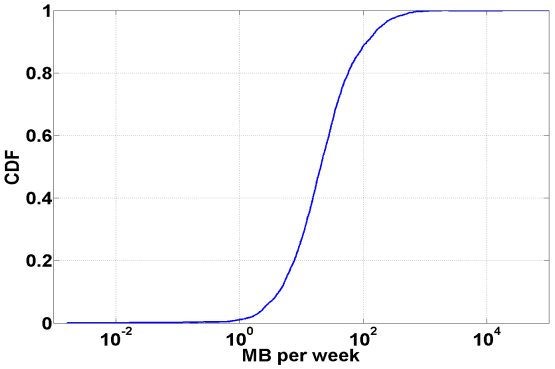


Figure 11: Feed rate CDF in MBytes per week. The median is roughly 20 Mbytes per week.

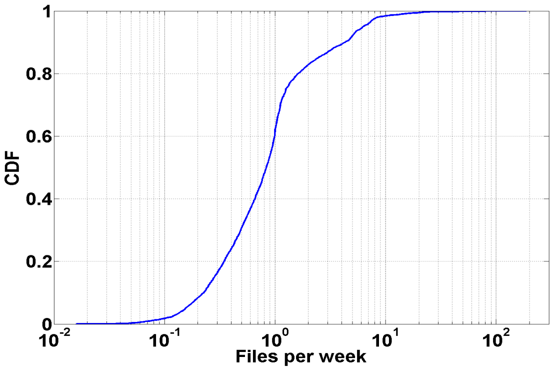


Figure 12: CDF of the number of files per feed per week. The median is around one file released per podcast per week.

the desired content. The maximum number of subscriptions per user is roughly 400.

A priori, it is not clear how a typical user would maintain his set of podcast subscriptions. Users may maintain a stable set of podcasts that they have subscribed and listen to over a long period of time; similarly, users may show a more dynamic behavior updating their subscription set regularly. Fig. 15 shows the distribution of the mean and median inter-subscription time per user. Fig. 15 indicates that subscriptions are batched over time with the median being significantly larger than the mean. While more than half of the subscription events are spaced at the timescale of days, the median value per user is in the order of minutes. We further examined the rate at which users update their subscriptions either by adding a new or removing an existing podcast from their subscription sets. The mean and median values per user are essentially similar to the distributions of Fig. 15.

The rate of adding new subscriptions is typically higher than removing the existing subscriptions. Fig. 16 presents this blow-up of subscriptions by presenting the difference of subscriptions and unsubscriptions per user, versus the number of total subscriptions on the *x-axis*. This suggests that, in scenarios of device-to-device dissemination, system designers may need to deploy automated mechanisms for prioritizing the content synchronization, rather than solely relying on the information about podcast subscriptions.

Table 3: File sizes in MBytes per mime type

Type	Median	Mean
Auddio	16.43	23.29
Video	29.39	65.40
Application	2.48	21.95
Other	20.41	38.90
All	17.70	30.54

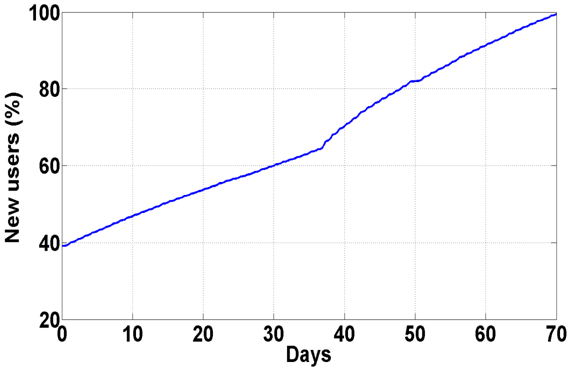


Figure 13: New users over time. The population increased by roughly 250% during the period of the study.

Information Gateways. What portion of users bring new content (i.e., podcasts) to the service? Such users essentially subscribe to podcasts for which no subscriptions exist before in our trace and may be regarded as *information gateways*; they are valuable in furnishing the service with references to new and potentially interesting content. Surprisingly, we find that only 1% of all distinct users introduced at least one podcast. Conditioning on these users, the portion of users who introduced n podcasts decreases roughly exponentially with n ranging from 1 to 5 podcasts. The decrease is significant with an order of magnitude decrease from n to $n+1$. These observations suggest that only a small portion of users introduce new podcasts and the majority of users exhibit a "follower" behavior by subscribing to existing podcasts in the service. Note that this is consistent with earlier findings in various online services and peer-to-peer systems, where, typically, only a small portion of users make most of the contributions (see, e.g. [6]).

4.2 Play Events

We now consider how typical users consume content, by examining play events. The distribution of the mean and median inter-play time of episodes per user is shown in Fig. 17. For half of the users, the median and mean inter-play times are less than 2 hours and 1.5 days respectively. The median value is significantly smaller than the mean value indicating existence of play sessions. This is rather natural as users appear to listen to a batch of episodes

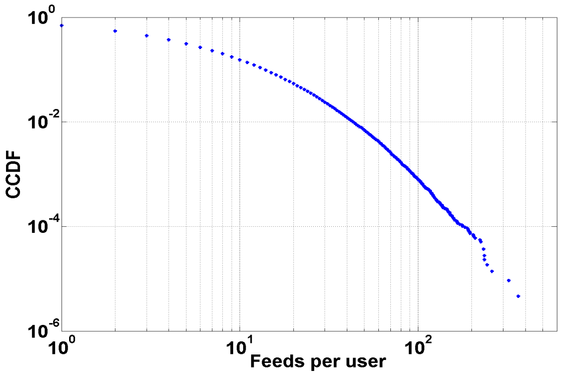


Figure 14: CCDF of subscriptions per user. The median is around three podcast subscriptions per user.

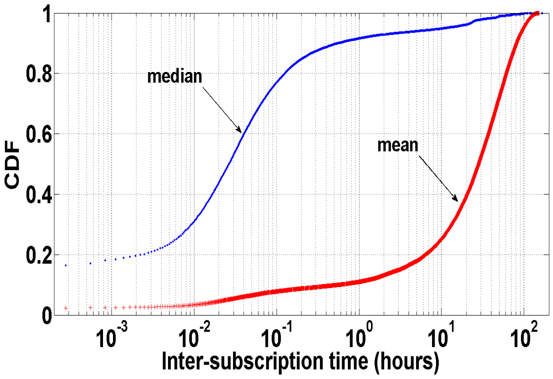


Figure 15: Inter-subscription time CCDF. The difference of the mean and median values suggests batched subscriptions over time.

in succession over time with small inter-play times, and such batches of episodes are separated with longer inter-play times.

Play events and devices. Play events may occur both at the Zune device and also at a personal computer. As such, one would expect that most users would play podcasts from either of the two types of devices (e.g., using the Zune device while away from home or office and personal computer otherwise). On the contrary, we find that most users play podcasts exclusively on one type of device. Concretely, we found that 36.55% of the users play podcasts exclusively from personal computers, 60.63% exclusively from Zune devices, and only a miniscule 2.82% of users play podcasts from both. In retrospect, this finding may appear natural as users may establish particular habits and stick to specific routines when interacting with the service. However, it is somewhat surprising that only a very tiny portion of users listen to podcasts using both device types.

Usage of particular devices, however, appears to correlate with time of day. Fig. 18 shows the cumulative fraction of plays across the time of day from either Zune devices or personal computers. While the respective rates follow each other over time, there is some noticeably larger rate of the use of Zune devices during working hours. Combining the previous observation, this finding may reflect a particular profile of users that listen to podcasts during working hours when away from their personal computers.

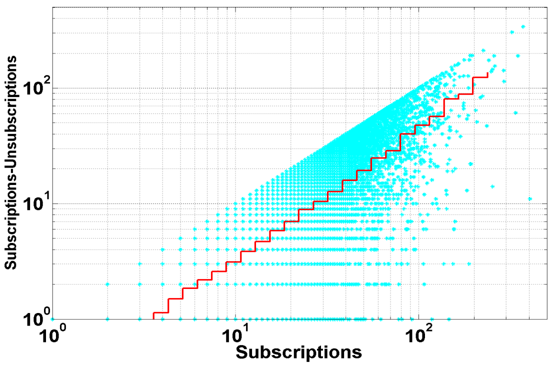


Figure 16: Subscriptions vs. difference of subscriptions and unsubscriptions per user. The line presents the mean across the *y-axis*. The subscription set per user appears to increase over time (i.e., the number of new subscriptions is larger than the one of unsubscriptions).

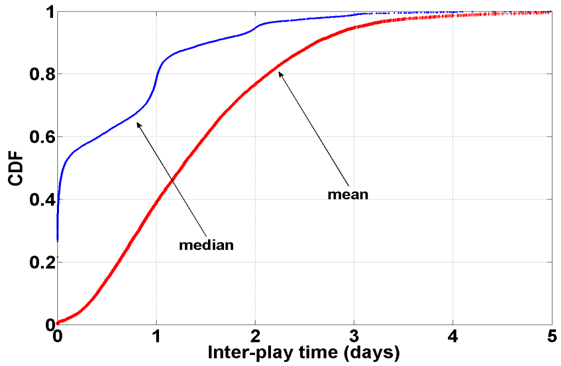


Figure 17: CDF of inter-play times for a given user. The difference between the CDFs suggests the existence of play sessions.

Subscriptions vs. play events. While subscriptions could be interpreted as explicit expressions of interest, they do not necessarily imply consumption of content. It is thus of interest to understand how this expression of interest correlates with actual consumption of the content, i.e. playing of episodes of the subscribed podcasts. This is informative to assess effective demands on the content synchronization and dissemination.

Fig. 19 shows the distribution of distinct podcasts played over time intervals of varied lengths that cover a range of timescales from minutes to weeks. Surprisingly, we observe that more than half of the users play no podcasts over time intervals of length up to a week. This is consistent with our earlier observation that the mean inter-play time of podcasts

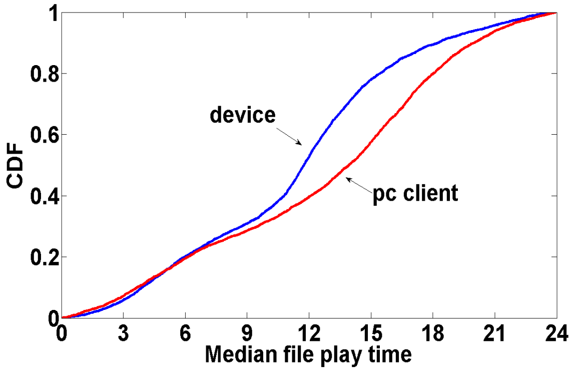


Figure 18: Time of day when files are played. Zune devices appear more frequently used during working hours.

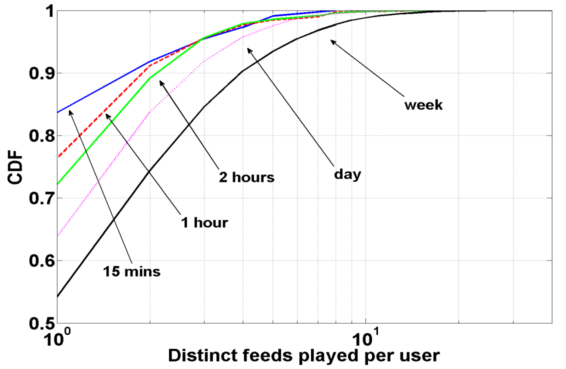


Figure 19: CDF of distinct feeds played per user over various time windows. More than half of the users play no podcasts over time intervals of length up to a week.

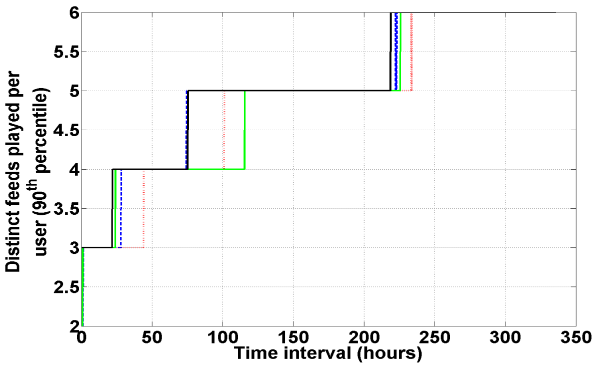


Figure 20: Number of distinct feeds played over a time interval vs. the interval length (90^{th} quantile). Distinct curves correspond to different observation instants. Typical users play less than 3-4 distinct podcasts.

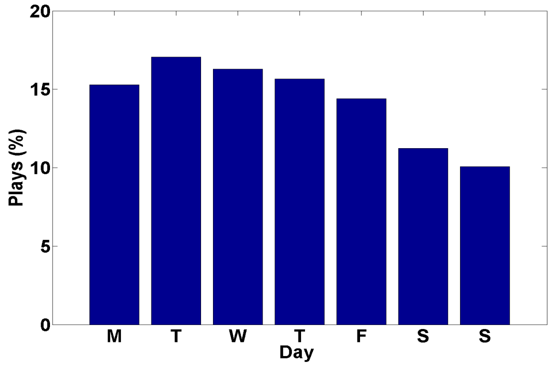


Figure 21: Percentage of plays vs. day of week. Most plays are during weekdays.

is 1.5 days over all users (see Fig. 17). In general, all but 10% of users play less than 4 distinct podcasts over time intervals of length of a week or less (less than 10 podcasts for all but 1% of users). These observations suggest that a typical user listens to a few distinct podcasts over time intervals shorther than a week. In order to better understand the dependence on the length of the observation interval, we consider the 0.9-quantile of the number of distinct podcasts played over users for a range of lengths of the observation interval (Fig. 20). We find that over day-long intervals, a typical user plays less than 3 to 4 distinct podcasts and this number is less than 6 over week-long intervals. Hence, while some users may be subscribed to a large number of podcasts, they typically listen to a few over a timescale of days. This observation could be exploited for the design of prioritization schemes for content synchronization and dissemination, where podcasts with high consumption probability would be synchronized first. With respect to weekly patterns, we observe that most play times occur during weekdays (Fig. 21), conforming with our observation that working hours observe larger play rates.

4.3 Popularity of Podcasts

The analysis of the popularity of podcasts is significant when considering peer-to-peer assisted or device-to-device dissemination of the podcasts. Fig. 22 displays the CCDF of the percentage of subscriptions per podcast. The distribution best fits to a log-normal distribution. Roughly 1% of podcasts have more than a few percentages of the total users. The CCDF of popularity with respect to the number of plays shows similar trends (Fig. 23).

To examine how podcast popularity compares across subscription and play events, Fig. 24 provides a first comparison of the podcasts' rankings based on user subscriptions versus the number of plays per podcasts. 80% of play events are generated by the 20% most subscribed podcasts, and as expected, the two rankings seem positively correlated.

To further illustrate the difference between the two rankings, Fig. 25 plots the fraction of the k most subscribed podcasts that exist also in the k most played podcast set, as k varies.[2] It is quite surprising to notice that only roughly

[2] Note that the respective metrics in Fig. 24 and Fig. 25 are in fact standard measures for comparison of rankings

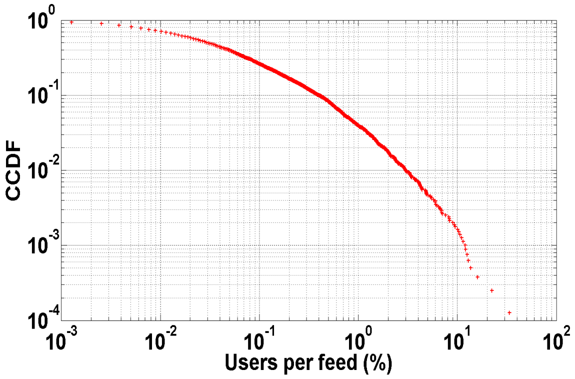


Figure 22: CCDF of users per feed. Only a few feeds (less than 3%) are subscribed to by more than 1% of the user population.

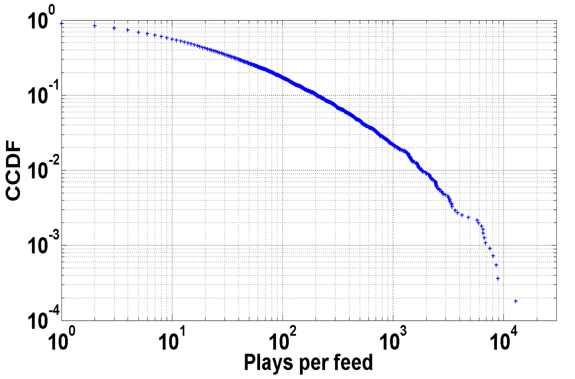


Figure 23: CCDF of plays per feed. Roughly 2% of the podcasts have been played more than 1000 times.

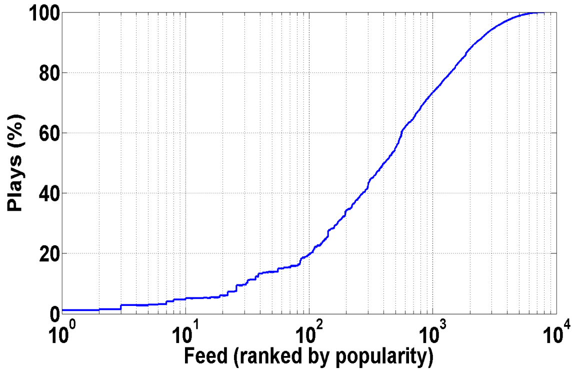


Figure 24: Plays vs. subscription rank. 80% of the plays are for the 20% most subscribed podcasts.

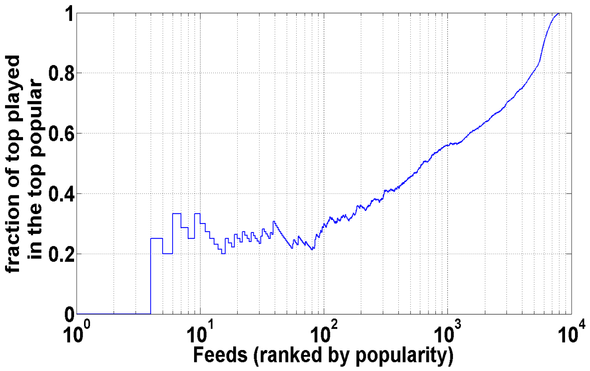


Figure 25: Fraction of top played feeds vs. subscription rank. Only around 20% of the most subscribed podcasts are within the 100 most played ones.

20% of the most subscribed podcasts are within the 100 most played podcasts. This might be an effect of the interface on the Zune Social site, where users are suggested a list of the most popular and most recent podcasts. Such podcasts then would receive subscriptions with higher probability, even if they are not the true most popular podcasts (if true popularity is based on the number of plays). Interestingly, we see no plays for 30% of podcasts throughout our dataset. Table 4 further highlights this difference of popularity in terms of subscribed versus played podcasts by displaying the set of the top-3 most popular podcasts in each case.

4.4 Popularity Reinforcement

User subscriptions to podcasts may be influenced by a number of factors. These factors may range from intrinsic interest for the content of the podcast to simply following other users by subscribing to popular podcasts. Lists of “what’s hot” or “what’s popular” are common in feed aggregator sites, and similar lists exist in Zune Social that typically displays a handful of the most popular podcasts (e.g., top-5). Additionally, users may browse podcasts through offered categorizations, thus conditioning podcasts under specific topics. Here, we attempt to provide some hints as to how the matching of users to podcasts occurs, by examining factors that could influence user choices. Our results suggest the existence of the reinforcement of podcast popularity, where “rich-get-richer” types of relationships prevail.

We first consider the subscription rate to a podcast given the number of subscriptions that the podcast has already received. Fig. 26 shows the average waiting time until the next subscription to a podcast that has already received n subscriptions. Except for small values of n (i.e., a few tens of subscriptions), the average waiting time seems to scale as $1/n$. Such a law could arise in the following hypothetical scenarios. For example, assume that the n users who subscribed to a given podcast are independent and have disconnected sets of friends to whom they can advertise the podcast; in such a case, we get the scaling in $1/n$ (if friends subscribe at a fixed rate). In another example, suppose that users tend to subscribe to a podcast with the probability proportional the the current number of subscribers of the podcast, each after some independent random delay. In both these cases, the average waiting time until a new subscriber is inversely proportional to the current number of subscribers. Similar observations hold when we consider the rate of playing

in information-retrieval, known as weighted recall at k and precision at k.

Table 4: Top-3 popular podcasts in terms of subscriptions and plays

Subscriptions	Plays
`www.fox.com/foxcast/data/collections/familyguy.xml`	`www.adamcarolla.com`
`bestofyoutube.podshow.com`	`www.theonion.com/content/radionews`
`www.hbo.com/podcasts/standup/podcast.xml`	`www.971freefm.com/pages/podcast/43.rss`

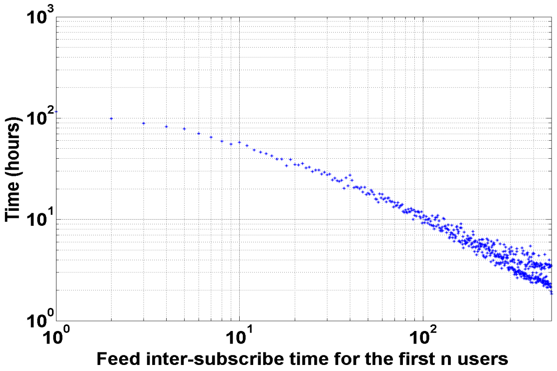


Figure 26: Subscription rate for podcast vs. the number of subscribers of this podcast. The time until a new subscription appears to scale as $1/n$.

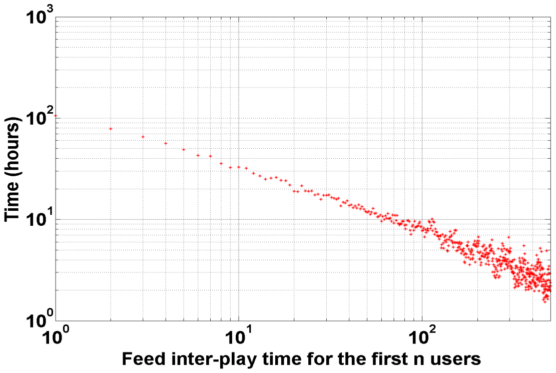


Figure 27: Play rate for podcast vs. the number of plays of this podcast. Similar scaling appears as in Fig. 26.

a podcast conditional on the number of times this podcast was already played (Fig. 27). As in the case of subscriptions, we observe that the same law holds.

We further estimate the probability that a user subscribes to a podcast with n subscriptions (Fig. 28). We find that this probability of selection is proportional to the number of subscribers n over a wide range of values from a few to thousands of subscriptions.

These observations suggest that the user choice in subscribing to and playing of podcasts is affected by their popularity. This may well be the result of exposing the popular podcasts more prominently in the user interface or a result of user-to-user recommendations. Overall, the probability of subscribing to a podcast increases linearly with the number of subscribers, which is consistent with standard preferential attachment models (see e.g. [7]).

5. PODCAST DISSEMINATION

In this section, we investigate various ways of disseminating podcast episodes to the users. Specifically, we consider three ways of downloading episodes from the Internet to user devices. The first natural way for users to download episodes is when the device is connected to the Internet, either at home through broadband connectivity, or via WLAN access points. The second way is to use the traditional 3G cellular networks. Finally, podcast episodes could be disseminated in an ad-hoc manner using device-to-device peer-to-peer assisted communication, i.e., users having podcast episodes in their buffer could act as relays and propagate these episodes to other users.

The choice of the dissemination method is guided by several factors, including the rate and the cost of the downloads, and the tolerance to delays of the users (e.g., are users willing to play podcast episodes just after they are published or can users wait to be directly connected to the Internet before playing episodes?). If the delay tolerance of users is high enough, say more than one day, the most economic and efficient way of downloading podcast episodes is obviously through direct Internet connection, e.g., at home. When the content of podcast is more time-critical, which could be the case for news or sports podcasts, users have to rely either on 3G cellular networks or on peer-to-peer assisted communication. It is worth noting that most of the 3G wireless providers already propose push-based services to broadcast news, e.g., live sport news. However, the volume of information broadcasted on these services is limited and much lower than the typical volume of podcast episodes. The cost of downloading podcast episodes through 3G networks may be significant, especially in roaming situations. Note that 3G providers usually impose limited monthly download cap, e.g., Sprint and AT&T typically impose a 5 GBytes monthly limit, and a few hundreds MBytes when roaming. For this reason, users may prefer to download podcast episodes through free peer-to-peer assisted systems rather than through 3G networks, especially when these users subscribe to several podcasts.

In the remaining of this section, we investigate the opportunity and the feasibility of peer-to-peer assisted podcast dissemination. Specifically, we first study user delay tolerance when playing podcast episodes; we investigate the typical volume of episodes that users have to download for example over a month; and finally, we examine the feasibility of device-to-device communication, e.g., are device-to-device contact times sufficiently long to transfer one or several typical podcast episodes?

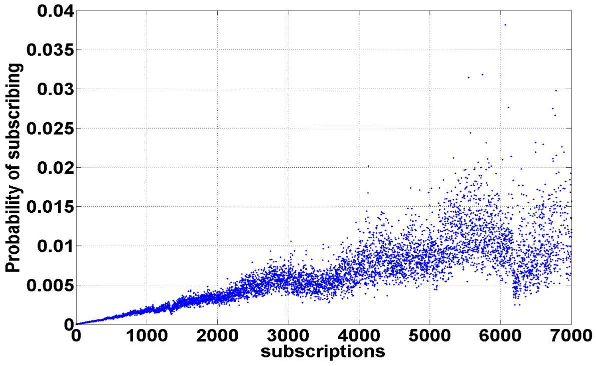


Figure 28: Subscriptions vs. popularity. The probability of subscribing to a podcast is proportional to the number of existing subscriptions of that podcast.

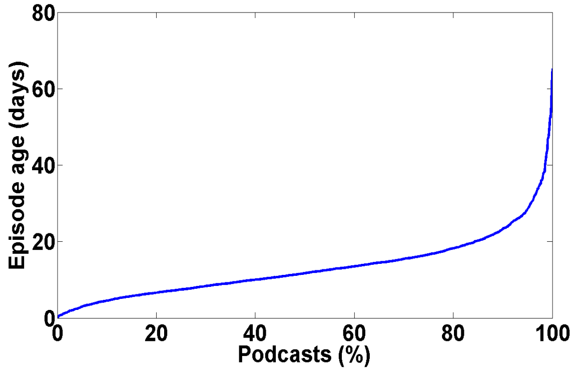


Figure 30: Median virtual age of episodes per podcast. Removing biases due to high-rate publish podcasts does not affect the median virtual age observed in Fig. 29.

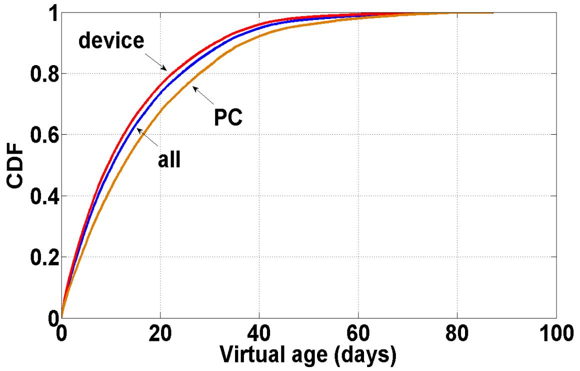


Figure 29: Virtual age of episodes. The median virtual age is around 10 days, which implies delay-tolerant users.

5.1 Delay Tolerance

It is important to know the time horizon over which the content is of interest to users since its publish time. In order to evaluate the horizon of interest, we consider the age of podcast episodes at their play times. Concretely, for each episode of a podcast, we record its publish time and the first play time of the given podcast by distinct users after a publish time. We call **virtual age**, the difference between the latter play and publish times. Note that the virtual age provides a lower bound on the actual age of the episode at its play time by a user, as the user may well play some other episode of the given podcast. (We had no information about which specific episode of a podcast is played by a user.)

The distribution of the virtual age is showed in Fig. 29. Note that this distribution is computed by taking samples of virtual age across all episodes and all users who played a released episode. Fig. 29 also shows the distribution of the virtual age, conditional on the type of the end-device (either mobile device or personal computer). Interestingly, we observe that the median virtual age is as large as about 10 days. Moreover, a significant portion of the episodes are of the age of a month or longer. We also note that playing on mobile devices typically occurs sooner after the publish time than on personal computers.

The distribution in Fig. 29 may be biased to podcasts with high episode publish rates and to popular podcasts whose episodes are played by many users as the distribution is for the aggregate of samples across episodes and users. In order to account for these biases, for each podcast we compute the median virtual age; the values are presented in Fig. 30. Again, we observe that the median of median virtual ages over podcasts is in the order of 10 days (11.73 days), thus of the same order as already observed in Fig. 29.

We further studied whether the episodes of the more popular podcasts are played sooner than those of the less popular ones. In Fig. 31, we show samples of the virtual age versus the total number of subscriptions of the corresponding podcast along with the average values computed over bins of the number of subscriptions. Interestingly, the results suggest a lack of bias with respect to the popularity of the podcasts; the mean virtual age remains roughly constant over a wide range from a few to tens of thousands of subscriptions.

We examined the same statistics as reported above under various other conditionings such as on the category of podcasts, individual podcasts, and the time of day and in all these cases the observations consistently remained; we omit the details for space reasons.

In summary, we observed that for typical podcasts, the delay tolerance is in the order of tens of days; we found that the median delay tolerance is about 10 days and is 1 day or shorter for as few as 1% of podcasts. These are important observations; they suggest that for typical podcasts, it suffices to synchronize the content a few times a day. Thus, it seems sufficient to rely on the dissemination through sporadic access to the Internet (e.g., while the device is connected to a PC when the user is at home or office). The P2P case would be of interest for the dissemination of podcasts whose delay tolerance is less than a day, case which was rare in our data. However, such scenarios might become prominent in the future as interactive-like services such as twitter grow in popularity. We note also that P2P dissemination would be of interest in scenarios where the Internet access is limited, e.g., while on travel or at places with limited infrastructure.

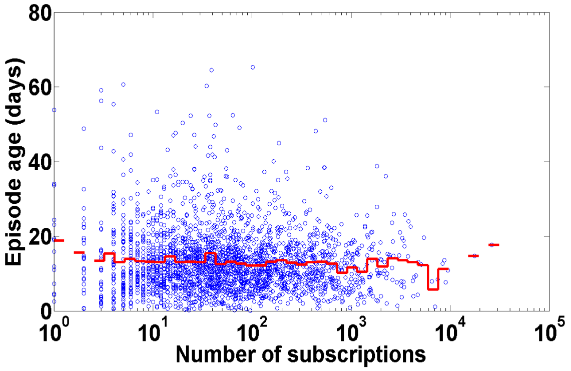


Figure 31: Virtual age of episode vs. podcast popularity. The virtual age does not appear correlated with the podcast popularity.

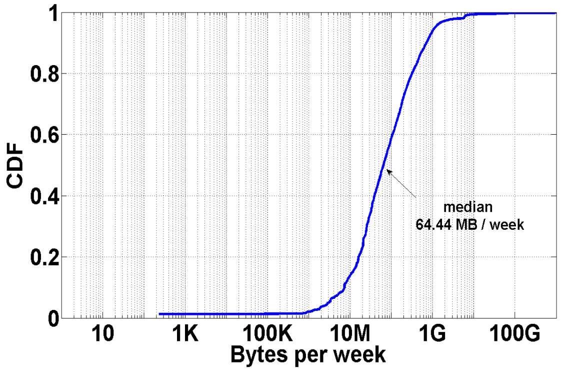


Figure 32: CCDF of rate of downloaded Bytes per week over users. The median is around 65Mbytes per week per user.

5.2 The Download Costs

Here, we evaluate the volume of the podcasting content that needs to be downloaded by a typical subscriber. This translates to download costs as many wireless service providers offer data plans with caps beyond which charging is volume based. We compute the download rate per user by summing the generation rates of podcast subscriptions. Fig. 32 shows the distribution of the download rate per user measured in bytes per week. We observe that the median download rate is about 65 MBytes per week. This corresponds to about 2 to 3 files since we already found that the typical file size is in the range of 20 to 30MBytes (Section 3). This is also confirmed in Fig. 33 where we show the distribution of the download rate in files per week. The median download rate amounts to about 1/4 GBytes per month and this volume is larger for a sizable portion of users; note that from Fig. 32 we have the download rate in the range 10 MBytes to 1 GByte per week for more than 80% of users. We conclude that podcasting can amount to a significant portion of a potential wireless data plan, and thus the transfer through 3G would need to be limited due to cost reasons.

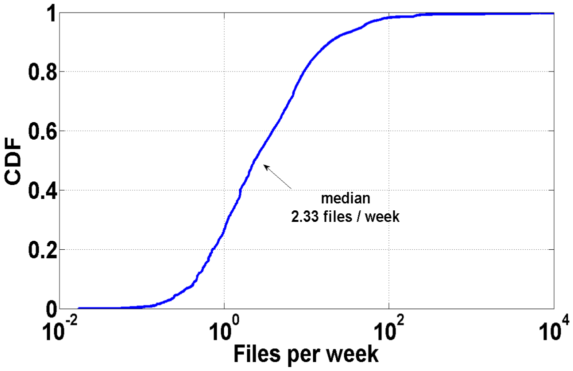


Figure 33: Number of files downloaded per week over users.

5.3 P2P Device-to-Device Dissemination

In this section, we consider the efficiency of P2P assisted dissemination that relies on device-to-device communications. In particular, we consider how the contacts between devices would be utilized under typical podcast file sizes and typical contact durations. For typical file sizes we refer to our analysis in Sec. 3 which suggests a typical file size of 15 MBytes for an audio podcast episode and 30 MBytes for a video podcast episode. For typical contact durations, we refer to the measurements results in the literature which suggest that many contacts may last 200 seconds or less; e.g., in [8], Fig. 8, we find that the contact durations are less than 100 seconds for about half of the contacts.

Table 5: Required contact durations for transfer of a 20 MByte file.

Transfer rate (Mb/s)	Contact duration (sec)
1 - 54 (WiFi)	160 - 2.93
100 - 480 (LAN/USB-2)	1.6 - 0.33
0.512 - 8 (WAN)	312 - 20

Table 5 shows us that under standard physical transfer rates, the time it takes to transfer a 20-Mbyte podcast file may well be in the order of a few hundred seconds. These transfer times would even be larger for a typical podcast video file. In view of the earlier discussion, the time to transfer a podcast file between two devices may well exceed typical contact durations. This suggests that one may need to resort to slicing of the files in chunks in order to improve the utilization the contact transfer opportunities.

6. RELATED WORK

Studies of podcasting services have been limited. Banerjee et al [3] were among the first to examine the characteristics of podcasts by downloading podcast content from 875 podcast streams. The authors examined properties such as the file sizes, the release times and proposed a model for file generation. Our study differs both on the set of properties examined and also regarding the scale of the examined podcasts.

Cha et al [9] performed a study of user generated content by crawling the YouTube and Daum sites. Their focus is on content popularity where the authors observed the presence of the Pareto principle. Our analysis confirms that this is also the case with podcast content. However, the overall focus of our work is on podcasting content where users subscribe to push-based services in contrast to the pull-based model of YouTube. The evolution of content popularity has also been examined in [10, 11]. The authors examine how the popularity of Flickr images evolves and study how information propagates through the Flickr social graph. Compared to these studies, in this work, we examine a broader set of properties pertinent also to device-to-device dissemination systems.

With respect to podcasting dissemination services for mobile devices, Lenders et al [12] were first to propose an architecture for a podcast dissemination system for mobile devices. Specifically, they proposed a P2P assisted dissemination and examined several device-to-device content dissemination strategies through simulations. Our work differs in that we base our study on real-world data of a podcasting service which we use to evaluate the actual benefits and feasibility of peer-to-peer assisted dissemination.

7. CONCLUSION

We have presented an exhaustive statistical analysis of current podcast services, by exploiting a 70-day trace from the Zune podcast social service. To stress-test the significance of the analysis, we have also partially analyzed additional podcast service providers, namely iTunes US and UK. To our knowledge, this paper reports the first study of large-scale podcast services both from publishers' and users' perspectives.

Specifically, we were able to characterize the statistical properties of podcasts, in terms of their publishers, their type and content, size and release frequencies of their episodes. We have also investigated the popularity of podcasts, based on both the number of received subscriptions and the number of times the corresponding episodes are played, and tried to infer how users subscribe to podcasts depending on their popularity. We further discovered that current podcasts are often consumed a long time after the release of their episodes, i.e., that the current podcast service is delay tolerant. As a consequence, efficient dissemination of podcast episodes is feasible even through a sporadic access to the Internet. This observation may be revised in the future as increasingly popular interactive podcasts appear.

8. REFERENCES

[1] Edison Research. Internet & Multimedia 2008 Reports . http://www.edisonresearch.com/home/archives/2008/04/internet_multim_5.php.

[2] eMarketer. Podcasting Goes Mainstream. http://www.emarketer.com/Article.aspx?R=1006937.

[3] Banerjee A., Faloutsos M., and Bhuyan L. N. Profiling Podcast-Based Content Distribution. In *IEEE Infocom Workshops*, pages 1–6, April 2008.

[4] V. Lenders, G. Karlsson, and M. May. Wireless Ad Hoc Podcasting. In *Sensor, Mesh and Ad Hoc Communications and Networks, 2007. SECON'07. 4th Annual IEEE Communications Society Conference on*, pages 273–283, 2007.

[5] Zune Social. http://social.zune.net/default.aspx.

[6] Jakob Nielsen. Participation Inequality: Encouraging More Users to Contribute, 2006. http://www.useit.com/alertbox/participation_inequality.html.

[7] M. Mitzenmacher. A Brief History of Generative Models for Power Law and Lognormal Distributions. *Internet Mathematics*, 1(2):226–251, 2004.

[8] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and C. Diot. Pocket Switched Networks: Real-world mobility and its consequences for opportunistic forwarding. Technical Report UCAM-CL-TR-617, Computer Lab, University of Cambridge, February 2005.

[9] M. Cha, H. Kwak, P. Rodriguez P., Y.-Y. Ahn, and S. Moon. I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System. In *ACM IMC '07, San Diego, CA, USA*, pages 1–14, 2007.

[10] M. Cha, A. Mislove, B. Adams, and K. Gummadi. Characterizing Social Cascades in Flickr. In *ACM SIGCOMM Workshop on Online Social Networks (WOSN)*, 2008.

[11] M. Cha, A. Mislove, and K. P. Gummadi. A Measurement-driven Analysis of Information Propagation in the Flickr Social Network. In *Proc. of the 18 Intl. World Wide Web Conference (WWW)*, 2009.

[12] Bychkovsky V., Hull B., Miu A. K., Balakrishnan H., and Madden S. A Measurement Study of Vehicular Internet Access Using In Situ Wi-Fi Networks. In *12th ACM MOBICOM Conf.*, Los Angeles, CA, September 2006.