Demystifying Porn 2.0: A Look into a Major Adult Video Streaming Website

Gareth Tyson
Queen Mary, University of London, UK
gareth.tyson@qmul.ac.uk

Yehia Elkhatib
Lancaster University, UK
yehia@comp.lancs.ac.uk

Nishanth Sastry
King's College London, UK
nishanth.sastry@kcl.ac.uk

Steve Uhlig
Queen Mary, University of London, UK
steve@eecs.qmul.ac.uk

ABSTRACT

The Internet has evolved into a huge video delivery infrastructure, with websites such as YouTube and Netflix appearing at the top of most traffic measurement studies. However, most traffic studies have largely kept silent about an area of the Internet that (even today) is poorly understood: adult media distribution. Whereas ten years ago, such services were provided primarily via peer-to-peer file sharing and bespoke websites, recently these have converged towards what is known as "Porn 2.0". These popular web portals allow users to upload, view, rate and comment videos for free. Despite this, we still lack even a basic understanding of how users interact with these services. This paper seeks to address this gap by performing the first large-scale measurement study of one of the most popular Porn 2.0 websites: YouPorn. We have repeatedly crawled the website to collect statistics about 183k videos, witnessing over 60 billion views. Through this, we offer the first characterisation of this type of corpus, highlighting the nature of YouPorn's repository. We also inspect the popularity of objects and how they relate to other features such as the categories to which they belong. We find evidence for a high level of flexibility in the interests of its user base, manifested in the extremely rapid decay of content popularity over time, as well as high susceptibility to browsing order. Using a small-scale user study, we validate some of our findings and explore the infrastructure design and management implications of our observations.

Categories and Subject Descriptors

C.2.0 [**Computer-Communication Networks**]: General—*Data Communications*

General Terms

Measurement

Keywords

Porn 2.0; Adult websites; Video streaming; Measurements

IMC'13, October 23–25, 2013, Barcelona, Spain.

http://dx.doi.org/10.1145/2504730.2504739 .

1. INTRODUCTION

The Internet has evolved from a largely web-oriented infrastructure to a massively distributed content delivery system [27]. Video content is particularly popular: by 2016, 86% of all Internet traffic is predicted to be video [5]. This transformation has led to a multitude of research attempts to characterise key video portals and the corresponding changing consumption patterns. This includes seminal studies into user-generated content (UGC) [13], video on demand (VoD) [40], Internet TV (IPTV) [14] and catch-up TV [6, 29]. Thanks to them, our knowledge has been expanded and, in many cases, the infrastructures improved.

However, there is an elephant in the room: adult video distribution. Similar to other kinds of content, adult video consumption has been undergoing dramatic shifts. Traditionally, adult videos were distributed via pay-per-view websites and within peer-to-peer communities (e.g., one estimate found that pornography constitutes up to 18% of the files on eDonkey in some regions [35]). Recently, however, increasing amounts of traffic are being generated by emerging YouTube-like websites that provide free on-demand access to adult videos. These sites term themselves "Porn 2.0", and give users the ability to upload, view, rank and comment on videos, as well as form online profiles.

Next to nothing is known about the nature of Porn 2.0, nor the way users interact with it; little is even known (outside ISPs) about the actual amount of traffic generated by these sites. Despite this, its prominence in the Alexa rankings [1] is undeniable, with six adult websites listed in the top 100, more than any other genre of video streaming. We confirm the huge scale of adult content in this paper, where we find over 111 million requests to a single adult website in just a three day period. Considering this scale and prominence, we believe it crucial to gain a better understanding of the characteristics of Porn 2.0, and derive principles that could mitigate its impact on the network.

In this paper, we inspect one of the most popular Porn 2.0 websites: YouPorn [2]. Founded in 2006, it has quickly risen to global prominence. For the last 5 years, it has been amongst the most highly ranked sites listed in Alexa, consistently appearing in the top 100. Due to its large scale, YouPorn provides an ideal case study of the Internet's expanding adult video market.

We repeatedly crawled the YouPorn website to extract information about the videos being created and watched. The entire corpus consists of 183k videos, consisting of footage spanning in excess of 3 years. Over their lifetime, these videos collected more than 60 billion views, confirming YouPorn's huge popularity. In the last 7

days of the traces alone, 912 new videos were added and viewed over 38 million times.

Using this data, we characterise the corpus, highlighting how users interact with various aspects of the system. Further, we make a number of observations that provide evidence of its specific properties compared to other types of media. Because our data does not provide information about users' personal intentions, we further augment our data with a small scale user study (46 participants) to reflect upon the reasons behind our findings.

One running theme in many of our findings is that despite being consciously constructed as an adult site comparable to YouTube, YouPorn differs from traditional UGC sites (e.g., YouTube, Vimeo) in two key ways. First, the number of user-driven content uploads on YouPorn is comparatively low. Indeed, YouPorn itself has uploaded 36% of the overall corpus. Second, despite the lower number of videos, each video gets many more views, on average, than more mainstream UGC sites. This smaller, more popular content corpus suggests lower operating costs for such websites.

Another theme is that users appear to be flexible about which videos they consume. Whilst we verify this independently using a small scale user study, our dataset reveals that this user flexibility manifests itself in two unexpected ways, both of which point to the importance of helping users to locate content. First, although adult video content is not expected to age with time, unlike temporally-sensitive genres such as news or weather, most of the views are garnered in the first days after upload. We demonstrate that browsing order is a strong factor that affects the number of views obtained: easy to find videos collect most views. Second, we find that the number of categories a video appears in strongly correlates with the number of views it obtains. Videos which are not categorised suffer severely. We also discover that no attempt is made by YouPorn at engineering the content of individual categories: in many cases, highly populated categories have too few videos.

A list of contributions and paper roadmap follows:

1. We offer, to the best of our knowledge, the first large-scale measurement study into so-called "Porn 2.0" adult video distribution on the Internet. We present our dataset in Section 3 consisting of over 60 billion views. We will make the data publicly available.

2. We provide a detailed analysis of key characteristics of adult video content (Section 4), as well as the way in which users interact with this type of corpus and its various categories (Sections 5 and 6).

3. We explore the reasons behind our findings through discussions fuelled by a user study (Section 7). Using this, we explore potential improvements that would benefit both network operators and content providers.

2. RELATED WORK

Pornography is anecdotally the most searched for content on the Internet. Many theories exist to explain this. Cooper [16] attributes this to the Internet's "triple-A-engine": Accessibility, Affordability, Anonymity. Suler [36] expands this into 6 factors, coined as the *online disinhibition effect*. Whereas, much work has gone into looking at who engages in online sexual activities and why they do so [12, 20], little is known about about the actual engines that underpin its distribution, especially the expanding Porn 2.0 phenomenon. This has seen websites emerging (e.g., YouPorn [2]) that allow users to upload, view, rate and comment on videos for free, much like YouTube does.

Name	Period	# Vids	# Views
Snapshot	28/02/2013	183,639	61 billion
3 Day	3/03/2013	183,591	111 million
Daily	1/03 – 4/05/2013	1656	96 million

Table 1: Overview of datasets.

A few studies have provided estimates of the demand handled by these websites. For example, Ogas and Gaddam [30] mention that Porn 2.0 sites such as Pornhub, RedTube, xHamster and YouPorn can gain up to 16 million views per month. This is a very conservative estimate compared to YouPorn's report of 100 million page views a day [3]. Other estimates have suggested that sites like YouPorn have a peak traffic rate of 800 Gbps [7]. Despite all this evidence, we still have quite a rudimentary understanding of the true scale of these services. Regardless, most experts agree that Porn 2.0 is a huge emerging economy that is not, as of yet, fully understood [17, 9, 21, 39, 30].

With this in mind, it is surprising to find next to nothing reported on the (systems-level) nature and operation of online adult multimedia delivery services. Instead, various research communities have focussed on specific sub-components such as automated recognition and classification [28, 25]; pornographic practices, communities and subcultures [9]; interest recommendations [34]; security issues [39]; and illegal content dissemination [26]. To our knowledge, this paper presents the first large-scale systems-perspective study of an online adult multimedia delivery service. We believe this work to be crucial, considering the increasing prominence of video distribution [5], of which adult media will likely continue to make up a significant proportion in the future [35]. That said, there are a multitude of studies into more traditional video streaming systems that already provide some insight. These include catch-up TV [6, 29], user generated content [13, 41], VoD [40] and IPTV [14, 22, 24]. Studies such as these have provided a range of insights, including content popularity models [23], optimised caching techniques [6] and improved delivery schemes [8]. As of yet, however, we are unaware as to how these principles apply to adult media systems. The rest of this paper therefore explores this topic.

Note that we do not make a sociological statement within our work; nor do we espouse the proliferation of pornographic media. We are interested in such media's impact on the network and, as such, we believe it is important to gain a better understanding of its characteristics.

3. YOUPORN DATASET

We crawled the YouPorn website to obtain information about its corpus and user base. Each video offered is accompanied by metadata that we extract. This metadata includes the number of views, the video rating, the number of ratings received, the upload date, the user who uploaded the video, the number of comments, and any categorical information. To collect this information, we performed a number of separate crawls, embodied within three datasets, summarised in Table 1.

Our first dataset, which we term the *snapshot* trace, contains information about all videos in the corpus $(183,639)$, collected on the 28/02/2013. We observe over 60 billion views of videos with durations collectively spanning in excess of 3 years. To augment this, we also collected a second dataset, which we term the *3 day* trace. To obtain it, we re-crawled the same videos 3 days later (3/03/2013). It contained 183,591 videos, 48 having been removed. Using the *3 day* trace, we calculated the evolution of all quantitative metadata including popularity.

The mentioned traces provide two "snapshots" of all videos on YouPorn. This, however, does not give much insight into the temporal properties of individual videos. We have therefore also performed smaller-scale periodic crawlings to collect a time series of snapshots. The third dataset, which we term *daily*, traced 2172 videos added between the period of 1/03/2013 – 4/05/2013. For each video added, we retrieved all metadata on a daily basis to study how it evolved, offering insights into the lifetime of each video. From the full set, we filtered the entries to leave only complete videos in which we had in excess of 21 days recorded. This left 1656 videos with an accumulated set of 96 million views.

4. CHARACTERISING CONTENT CORPUS

In this section, we investigate the content corpus offered by YouPorn. We traced 183,639 videos within the corpus dating from September 2006. This constitutes their entire video repository, as available at the time of writing.

4.1 Content duration

First, we inspect the duration of videos within YouPorn, presented as a cumulative distribution function (CDF) in Figure 1.[1] We observe from Figure 1 that most videos are rather short. About 80% are shorter than 15 minutes, with a very small fraction of them exceeding 45 minutes. If we divide them into 1 minute time ranges, the largest bucket is the 5–6 minutes, which contains 25% of all videos. This propensity could have emerged for a number of reasons. An obvious one is the presence of commercial videos that are intended to advertise content from other (e.g., pay-based) websites. Such users tend to upload short previews of longer videos in an attempt to entice users to their websites. YouPorn allows banners to be placed below videos to better enable this. It also appears that many other videos only contain relatively short scenes, without the sorts of preambles seen in other media types. Practically speaking, the corpus therefore appears very much like a convenient "pick and mix" repository where users can select snippets of videos that suit their interests rather than watching entire films. Whereas the reasons for this style of viewing could be diverse, it is important to note that Alexa reports the average viewing time on YouPorn as only ≈9 minutes [1]. With such time limitations, uploaders (particularly commercial ones) must ensure that only the most interesting elements of their films are seen by viewers. Lastly, Figure 1 also shows the duration of videos weighted by the number of times they are watched. The curves are near identical, indicating that users do not have a particular preference for one duration but, rather, watch various durations equally often.

4.2 Content injection

An extremely important component of Porn 2.0 is the injection of content by users. We therefore inspect the frequency at which videos are uploaded into the corpus. Indeed, the short durations of the videos could come with significant churn to sustain the interest of the user base. Figure 2 first provides a CDF of the number of daily video uploads. On average, only 78 videos are added per day, a surprisingly modest figure compared to sites such as YouTube [4]. This low daily injection rate therefore suggests that rapid corpus expansion is not necessarily vital for the success of this platform. As a comparison, already in 2008, YouTube was reported to have well over 140 million videos — over 700 times more than YouPorn's current corpus of 183,639 videos. Despite this, according to Alexa, the number of pageviews for YouTube is just 100 times more than for YouPorn. This suggests that sheer volume of content is not necessary for the success of adult video streaming services. That said, Figure 2 does highlight that the number of daily uploads has increased notably over time, with an average rate of about 140 over the last 3 months of the trace.

[1]Some videos had bogus length fields (e.g., 1000 hours). Consequently, we manually removed all entries above 3 hours (74 videos), leaving 99.99% of the videos in the trace.

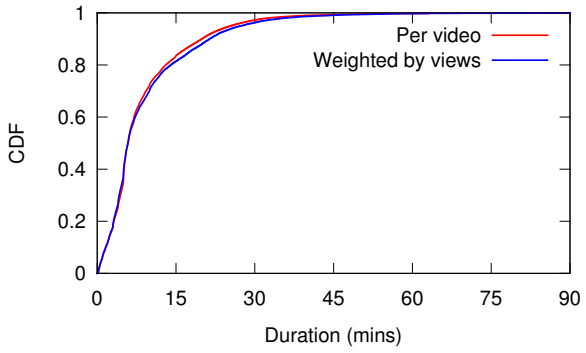

Figure 1: CDF of content duration.

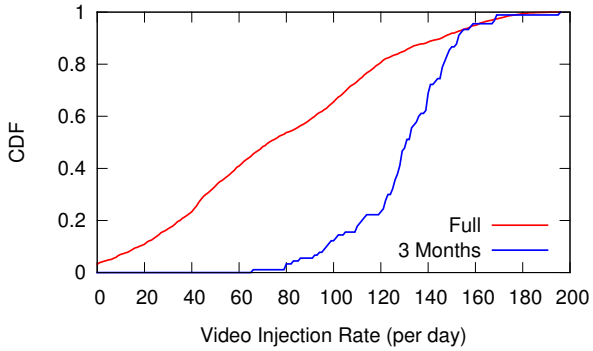

Figure 2: CDF of daily content uploads.

Beyond these absolute figures, we also examine *who* uploads content. We find that only 5,849 distinct usernames have ever uploaded over the entire 6 year history of YouPorn. Note this includes "Unknown" users with anonymous uploads (33k videos). Furthermore, as indicated in Figure 3, most users (56%) upload only a single video, with the majority (80%) uploading at most 5. Figure 3 also presents the number of submissions per user in the "am-

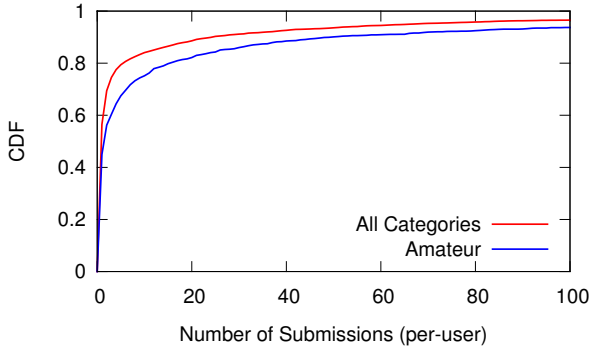

Figure 3: CDF of per user video uploads.

ateur" category, which one would initially imagine to have a far more proactive user-base. Even this category, however, shows very few uploads. In fact, overall, we observe 75 days during which no content was uploaded whatsoever.

This observation led us to perform manual inspection to better understand the nature of the uploads. We found that many uploads were actually provided by commercial producers. This observation extends to all categories, even "amateur". Figure 4 presents the daily upload rates over the entirety of YouPorn's existence. An upwards trend can be seen, suggesting an expanding base of uploaders. Investigation of these uploaders shows that 36% of the content is actually uploaded by YouPorn itself; a process that started almost 2 years after YouPorn's inception, with 39 videos, on average, being injected each day. These are all professional videos that are typically produced by a listed production studio. We conjecture that this may have been initiated, in part, to ensure a sufficient number of daily uploads. Regardless of the underlying reason, after YouPorn started uploading content, we observe that every day has new uploads, showing that YouPorn's own contributions have had a significant impact. In fact, without these contributions, the overall average daily upload rates would drop massively from 78 to 50 video per day. This can be seen in Figure 4 with extremely predictable and sustained upload rates boosting the overall uploads after year 2. Our observations therefore suggest that YouPorn is closer to a commercial platform than a user generated one.

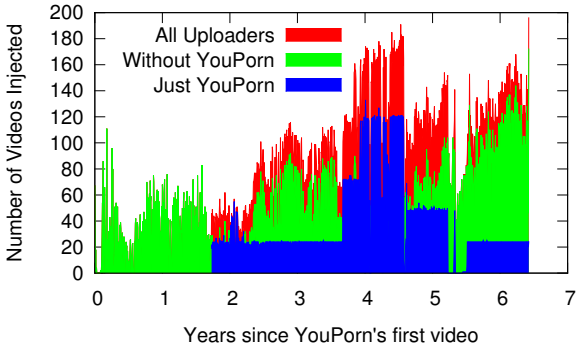

Figure 4: Breakdown for daily Upload rates (with and without YouPorn's contributions): During the first two years of YouPorn's existence, there were days without any new uploads. After this YouPorn itself has been uploading new videos, leading to a notable increase in the number of new uploads, suggesting that much of the content is not user-generated.

4.3 Content removal

So far, we have seen the number of videos added to YouPorn each day. An equally important aspect is how many videos are removed. Unlike most user generated content sites, YouPorn does not offer a straightforward way for users to remove their own content after upload. Instead, all removals must be requested — measuring removal levels therefore provides strong insight into the amount of content that deviates from YouPorn's policies (e.g., copyright issues).

YouPorn allocates each video a unique numerical identifier. These are selected from an incrementing pool of time-dependent identifiers. We surveyed a large range of the identifier space to collect the status of each video therein. We incrementally crawled all video identifiers between $7,692,093$ and $8,300,674$ in 1K blocks; this range covers March 2012–2013. Each identifier returns a status page, allowing us to ascertain the current status of each video upload.

Figure 5 shows the number of removals we observed across the measured identifier space. On average, we found that 11.7% of the content is removed. A number of notable spikes can also be seen; for example, we found that all videos were removed from a specific 1k identifier block. Manual inspection revealed many videos with production studios in their titles, suggesting possible copyright issues.

From this, one might assume that the majority of videos become active in the repository. However, we discovered other possible video statuses beyond "active" and "removed". We found a large number of videos that were classified as being "processed"[2]. This state is allocated to a video during the initial stages of its life when it is being encoded. It is therefore curious as to why many videos do not proceed beyond this state; on average, 61% of videos are still being processed even after several months of existence. We strongly suspect that some sort of (potentially manual) vetting procedure takes place. Consequently, only a minority of uploads are actually accepted for publication on YouPorn: only 18% of videos are active in the identifier range we studied. This also offers some explanation as to why the injection rates in YouPorn (particularly historic ones) are lower than could be expected for a repository of its prominence. Further, the need to vet content might also offer insight into why YouPorn started to upload a large number of its own videos.

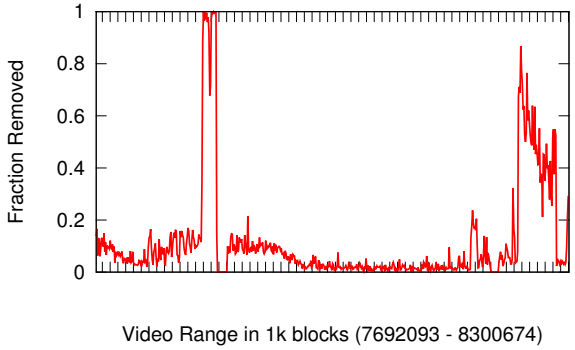

Figure 5: Removal rate of content.

4.4 Summary

We conclude that the users of such services do not particularly require huge novelty from the content available to them. It seems that a small number of new videos will still satisfy the demand. However, from the daily video uploads, it seems that the current user base requires new content to be available every day, forcing YouPorn to ensure a steady flow. Amongst the uploaded videos, only a limited fraction become available eventually with $\approx$ 10% removed.

5. CONTENT POPULARITY

We have seen that YouPorn is a constantly expanding repository, with new (typically short) videos being uploaded on a daily basis. Next, it is important to understand the way in which users interact with this corpus. Particularly, we are interested in seeing the pop-

[2]We also note 8.6% of videos in other miscellaneous states, i.e., "failed" or "not available".

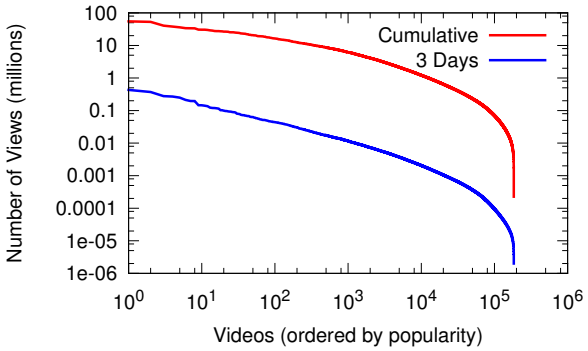

Figure 6: Number of views per video (log-log).

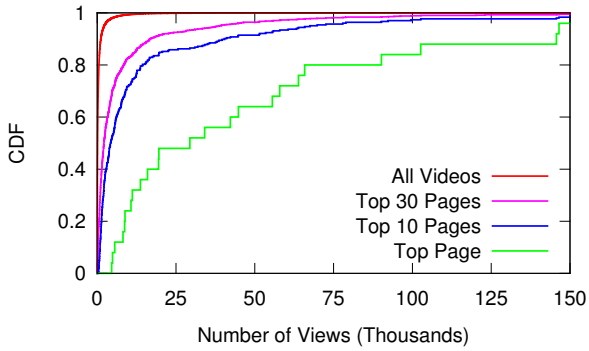

Figure 7: CDF of number of views per video.

ularity of individual objects, as indicated by the number of views they receive, and their respective importance in YouPorn.

5.1 Popularity: I'll take anything you've got

We begin by looking at the popularity of all videos taken within our traces. We rely on two different time windows: (*i*) the entirety of YouPorn's existence (cumulative) and (*ii*) a three day period.

One recurrent property that has been observed across many types of content repositories is a Zipf-like popularity distribution, evidenced by a straight line on a log-log plot. Figure 6 presents the number of views per video, on a log-log plot (ordered by rank) for both time windows. We observe a distinctive popularity skew, but not a straight line as one would expect from a Zipf distribution.

We make two complementary observations from Figure 6. First, the skew towards the "head" (or popular part) of the corpus is far less than has been previously observed in other UGC corpora. Specifically, the top 10% of YouPorn videos receive only 65% of the views. In comparison, the top 10% of videos generate 80% of views on YouTube [13], and 82% of views on Vimeo [33]. Second, the "tail" (or unpopular part) are correspondingly more popular. Nearly 93% of YouPorn's videos receive at least 10k views over their lifetime when inspecting the *snapshot* trace. In comparison, only 1.9% of videos on Vimeo generate more than 10k views. Further into the "tail", all videos in the YouPorn catalogue have at least 226 views, whereas fewer than 47% of Vimeo videos have at least 200 views.

Two explanations are possible. One possibility is that videos uploaded to YouPorn are generally of a higher quality than on other UGC sites; indeed, manual inspection reveals a wealth of professionally produced content. As such, a higher quality could encourage users to view a more diverse body of content. A second possibility is that users have a greater flexibility in their content selection requirements, i.e., users are not particularly selective in what they choose to watch, thereby resulting in views being more evenly distributed. Our user study (Section 7) suggests that the latter may well be true, with many users having far looser interest constraints than traditionally understood. For example, in mainstream VoD services, users often have relatively tight constraints on what they wish to watch. This might be a certain programme, a serialised TV show, or a particular genre [29]. Without these constraints, however, much larger sets of objects become acceptable for consumption, leading to a lesser skew in the popularity.

That said, flexibility does not explain the skew - if content selection were entirely flexible, why do some objects gain more views? One cause could be the way viewers discover content to watch. For instance, users of other user-generated video corpora such as YouTube may have particular videos in mind, driven by URL links from other websites, such as social networks. It is believed that up to 45% of requests to YouTube come from social sources [38]. Similarly, Borghol *et al.* [10] found that an uploader's social network (on YouTube) is one of the strongest predictors of video popularity. In contrast, users are less likely to share YouPorn links on social networks such as Facebook, or even discuss specific videos with friends. We conjecture that a lack of external referrals from other websites helps create an information bottleneck that prevents users from discovering the exact URLs corresponding to individual videos, thereby forcing most viewers to find videos through YouPorn's built-in facilities (like browsing or search). When combined with the inherent user flexibility, this likely predisposes any "generic" user to retrieve content from the easiest source possible, e.g., front-page listings.

To verify the above assertions, we correlate the number of views a video receives with the default front page browsing order. Figure 7 presents the outputs of this analysis for the *3 day* trace. We observe that the majority of views do, indeed, come from easy to access items. On average, videos on the front page[3] achieve 55k views, compared to an average of 9k for the top 30 pages. These can then be both contrasted with the overall average of just 603 views per video.

Our observations reflect well the type of behaviour one would expect from such a content repository. With a corpus in which it is difficult to differentiate objects, it is likely that only the most dedicated viewers (e.g., ones with special interests) would take the effort to find particular items of interest. More generic viewers seek easy access content, which, of course, creates a certain level of skew because all users are presented with the *same* easy access content items. However, due to the churn of the content, these objects are quickly pushed from the front pages, thereby flattening the popularity distribution into the one shown in Figure 6.

5.2 Popularity: But now I've changed my mind

Next, we look at how video popularity evolves over time, driven by regular content injection (Section 4.2) and what appears to be a largely flexible and browsing-driven user base (Section 5.1).

To gain insight into how videos accumulate views over time, we look at the distribution of views based on a content item's age. Figure 8 presents a log-log plot of the number of views per video, ranked by popularity. Each curve shows the distribution of videos with a given age (note that the tails are different due to a varying number of videos being uploaded on those individual days). We

[3]This is a conservative estimate as we do not include "featured" videos, which receive a more prominent status on the front page.

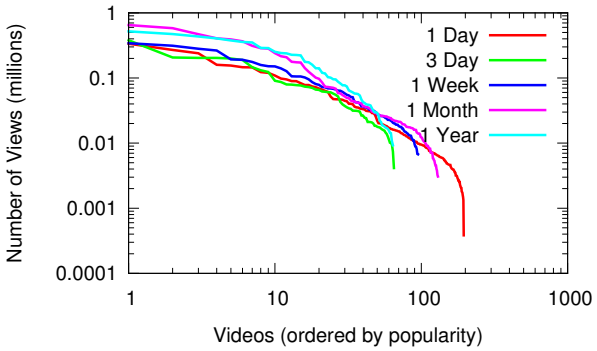

Figure 8: Number of views per video (log-log), for different time windows.

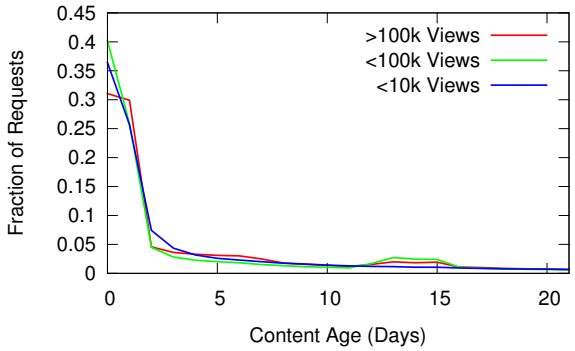

Figure 9: Evolution of number of views over time (per video).

find that videos uploaded a long time (e.g., a year) ago have not received particularly more views than recently uploaded ones. This is in stark contrast to prior UGC studies (e.g., [13, 37]) that show far greater cumulative views for older videos.

To explore this, we inspect the *3 days traces* to ascertain the most popular content ages during this short period. There is a distinct preference for recently uploaded videos. Content that was uploaded on the same day as the *snapshot* trace had collected on average 28k views, in contrast to an average of only 584 views for all other content ages. That said, we find notable exceptions: the content age with the third highest average number of views is 6 years, suggesting that there is no inherent reason why older content would not be suitable for viewing today. Note that browsing options (e.g., "most viewed") make such videos easy to find.

To understand how the characteristics of particular videos affect this rapid aging process, we inspect the *daily* traces, which show how the popularity of individual videos evolve over a more extended period of time. First, we partition videos into popularity groups based on the number of views they receive during their first day. We then average the number of views per day received by each video, and normalise that as a fraction of their total view count. Figure 9 presents the results. We observe a sharp decline in the number of views per day across all popularity groups, with the biggest decrease occurring after the second day.

We now draw a conclusion. Continuing from the discussions in Section 5.1, we see that the previously discussed user flexibility and browsing behaviour has a direct (and perhaps damaging) effect on temporal trends. YouPorn displays content on its front page in order of upload date (and then rating), thereby making more recent content easiest to find. Accessing content older than one day requires browsing through ≈4–5 pages of listings, a process which many users may find cumbersome.

Only the most popular videos ($>$ 100k views) can resist this decay, with similar viewing figures being recorded on the first and second day.[4] After the third day of their publication, however, even extremely popular videos are likely to be pushed down by ≈10 pages, making it significantly harder for users to discover them. Thus, on the third day, they immediately begin to exhibit the traits of their less popular counterparts, as their viewing figures plummet. Future views are then limited to users who are prepared to more proactively seek content of interest. Relating this to traditional UGC, Crane and Sornette [18] provide a classification of video types (memoryless, viral, junk and quality). This would place typical YouPorn videos into the category of junk. These are videos that experience a short burst of activity, followed by a popularity collapse. In comparison, such videos belong to the smallest category of YouTube videos [18, 32].

Therefore, YouPorn videos seem to have developed temporal properties similar to news and weather shows [6], which are highly temporally dependent (e.g., weather forecasts from the previous day are rarely of interest). However, contrary to news and weather content, there is no temporal dependency in adult content — videos created a year ago would still seem to be suitable for viewing today. Instead, it seems that this behaviour is formed from user flexibility: many videos meet the content consumption requirements of most users and, hence, are readily satisfied with the age-based listing on the front page.

5.3 Summary

We find that video popularity in YouPorn follows a far less skewed distribution than traditionally understood. The reason lies in the way users discover content, relying heavily on front page browsing. Due to the non-interest-specific nature of this default browsing, we conclude that most users have quite loose interest constraints, allowing them to be satisfied by a potentially large portion of the corpus. This results in a rapid decline in the number of page views; most videos die out quickly as they get pushed down the browsing order by newer published items. Consequently, we deduce that the level of skew observed is actually largely an artifact of the way content is presented to users, rather than any inherent aspect of the video content itself.

6. CATEGORY ANALYSIS

The previous sections have looked at the corpus as a single collection. However, videos within the corpus can also be listed under one or more category pages. Although these categories do not offer definitive information on the semantic nature of the videos, they allow us to inspect more targeted groups of content and their role in helping users to discover content of interest.

6.1 Category primer

In total, 62 categories[5] are available on YouPorn, spanning a range of interests. For each video, an initial category is chosen at upload time. Videos can also later be categorised further by other users once the content is published. Whilst this community-driven nature of categories means that some videos could be incorrectly

[4]Highly rated videos will appear at the top of the browsing list for the previous day.

[5]In fact, 63 exist but one was not populated; it had only been created a few days before the crawl and therefore we excluded it.

classified, category-specific web pages offer an additional mechanism by which users can browse content of a particular type; a mechanism that many users seem to find helpful (Section 7). We emphasise that categories on YouPorn are not free-form tags similar to folksonomies found on other Web 2.0 websites. In contrast to folksonomies, user choice is restricted to the well defined terms allowed by YouPorn.

We first inspect the number of videos in each category, shown in Figure 10. Due to their explicit nature, the categories have been pseudo-anonymised using their first two letters.[6] We observe a significant skew: the top 20% of genres contain 57% of the videos, whilst the bottom 20% contains less than 2%. In fact, only 11% of content belongs to the bottom 50%. YouPorn's categories therefore consist of a much longer tail than traditionally seen in UGC services [15], with far more categories for users to choose from.

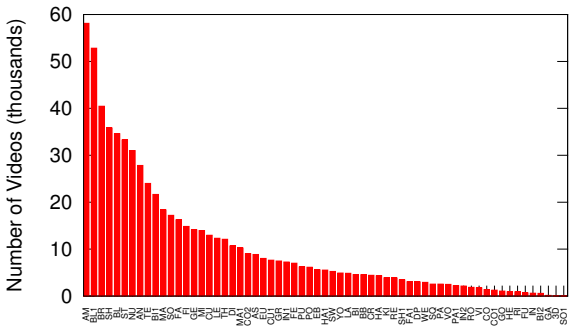

Figure 10: Number of videos per category (ordered by number of videos in the category).

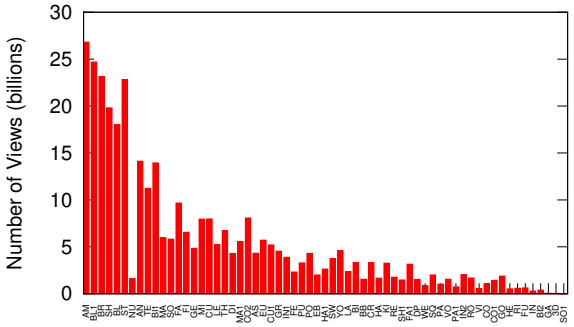

Figure 11: Number of views per Category (ordered by number of videos in the category).

We also inspect the collective popularity of these various categories. Figure 11 presents the accumulated views for each category. We observe a trend similar to Figure 10, with well populated categories receiving many views. The most notable exception is "NU" (videos without a category), which does extremely poorly in relation to its size in the corpus. Whereas 17% of the corpus has no category, these videos only collect 2% of the views. Table 2 provides an overview of the top 20 categories, ranked by size.

6.2 Efficiency of categories

We have seen that categories play an extremely important role in discovering content in YouPorn, and that not participating in this process has dire consequences in terms of views for uncategorised videos. However, categories offer insight into the interests of both uploaders and users, as well as their relationship. In a sense, Porn 2.0 sites could be considered as a form of marketplace where uploaders present their videos for consumption, competing for audience views. One should therefore strive for a marketplace in which the supply for a category exactly matches its demand. We capture this principle through the concept of corpus *efficiency* (or, more accurately, *inefficiency*). An efficient corpus is one in which the fraction of views for a category exactly matches the fraction of the corpus that the category constitutes. We measure the inefficiency, $\mathcal{I}$, for each category as:

$$\mathcal{I} = \begin{cases} \frac{V}{C} - 1, & \text{if } V > C. \\ -(\frac{C}{V}) + 1, & \text{otherwise.} \end{cases} \quad (1)$$

where V is the fraction of views that the category receives, and C is the fraction of the corpus that the category constitutes. As a video can have multiple categories tagged to it, we utilise two ways of calculating these fractions. The first approach,[7] termed *inefficiency*, attributes one view to each category that the video is part of; consequently, a video with two categories, "BL" and "AM", will have one view allocated to each. Obviously, this first approach will also artificially inflate the corpus size. The second approach, which we term *weighted inefficiency*, splits the views of a video equally between all categories it belongs to. Specifically, the number of the video's views attributed to that category are factored by $\frac{1}{\kappa}$, where κ represents the number of categories a video has. In this case, the number of videos attributed to the category are similarly factored down by $\frac{1}{\kappa}$.

For both weighted and unweighted inefficiency, if the value of $\mathcal{I}$ is above 0, it means that a category receives a disproportionately large share of the views, whilst a value below 0 indicates that a category receives disproportionately fewer views than would be

	Videos	Views (bn)	Inefficiency	Colocation
AM	58114	26.82	-0.11	5.69
BL1	52814	24.69	-0.10	6.37
BR	40447	23.14	0.12	6.79
SH	35815	19.81	0.08	7.44
BL	34607	18.04	0.02	6.67
ST	33259	22.81	0.34	7.10
NU	31004	1.59	-8.97	1.00
AN	27832	14.13	-0.01	5.65
TE	23999	11.24	-0.09	6.61
BI1	21667	13.94	0.26	6.57
MA	18348	5.96	-0.58	5.74
SO	17204	5.81	-0.52	5.63
FA	16240	9.66	0.16	7.64
FI	14844	6.55	-0.16	6.83
GE	14093	4.84	-0.49	5.44
MI	13944	7.96	0.11	6.74
CU	12959	7.98	0.20	7.46
LE	12315	5.21	-0.21	4.67
TH	12097	6.71	0.08	6.07
DI	10660	4.29	-0.27	6.33

Table 2: Category rankings ordered by number of videos. Categories appearing in other top 20 rankings (e.g. inefficiency) are not necessarily captured in this table.

[6]We make mappings available online at `http://www.eecs.qmul.ac.uk/~tysong/yp/mappings.txt`

[7]Unless otherwise stated, this calculation of inefficiency is used throughout the rest of this section.

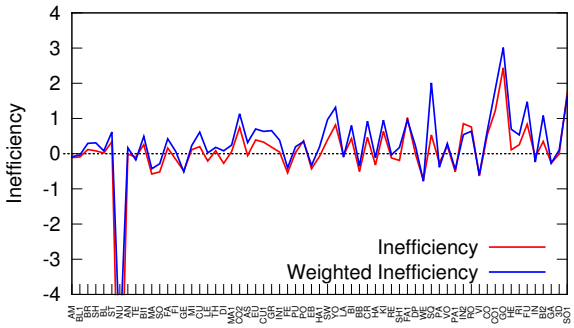

Figure 12: Inefficiency value for each category (ordered by number of videos). For null, $\mathcal{I} = -8.96$ and weighted $\mathcal{I} = -5.51$ (this is cut-off to improve readability).

expected. More generally, a value above 0 represents a category in which "demand outweighs supply" (popular), and a value below 0 represents a category in which "supply outstrips demand" (unpopular). This provides a normalised view of popularity, in contrast to the absolute one given in Figure 11.

Figure 12 presents the inefficiency levels for all categories, with both measures of inefficiency showing very similar trends. First, we observe that using the absolute number of views a category receives is somewhat misleading. The category with the highest viewing figures ("AM") actually receives fewer views than could be expected from its size in the corpus ($\mathcal{I} = -0.11$). It seems that this category collects views through its dominance in the corpus, rather than through an excessive demand for the genre. This lack of efficiency is observable in all other categories too — several unexpected genres have a disproportionately large number of views, whereas other genres have too many videos and too few views. 27 categories have a disproportionately large number of videos in comparison to the views received ($\mathcal{I} < 0$), whereas 35 categories are disproportionately popular compared to their size in the corpus ($\mathcal{I} > 0$).[8] No categories were found to be truly market efficient with "BL" coming closest at 0.016, alongside 14 others that fall between -0.1 and 0.1.

6.3 Category Colocation

The previous subsection has shown that there are some notable market inefficiencies in YouPorn's corpus. These inefficiencies could offer a significant opportunity for uploaders. For example, theorists believe that users constantly seek out new forms of visual stimulation [31]. Therefore, some less populated categories would be suitable for targeted content injections as their demand outstrips supply.

However, the ability to allocate a video to multiple categories (i.e., category colocation), could undermine the independence of the samples — a video in multiple categories will be far more visible through category-based browsing. To investigate this, Figure 13 re-plots Figure 12 whilst also presenting the average number of colocations for each category. For example, Figure 13 shows that "MA1" is, on average, tagged in a video alongside 6 other categories. We observe a strong correlation between inefficiency[9] and colocation, with a correlation coefficient of 0.66. For instance, the "GO" category gets 3 times as many views as could be expected from its proportion of the corpus. However, it is by far the most colocated category. On average, videos categorised as "GO" are also placed in 10.36 other categories. We find that 4 out of the top 10 categories, ranked by the colocation level, are also in the top 10 ranked by normalised popularity. This confirms our suspicion (backed by the small-scale user study; see Section 7) that category-based browsing is an intensively used tool.

[8]Interestingly, many of the more unusual niche categories (e.g. "GO", "IN2", "FA1") fall in this area.

[9]Note, inefficiency also offers a measure of normalised popularity.

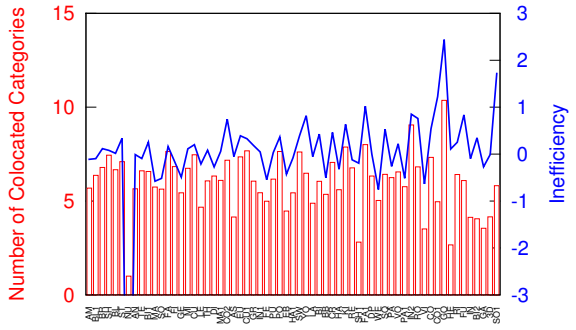

Figure 13: Number of colocated categories (ordered by number of videos in the category).

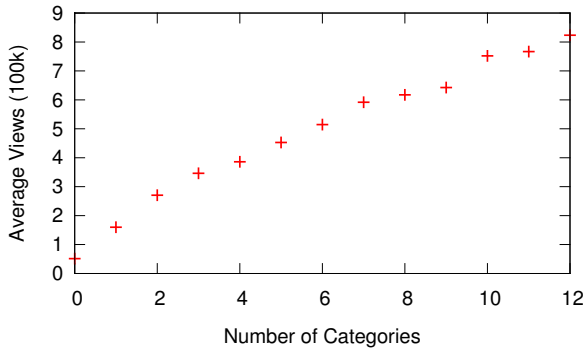

Figure 14: Number of views correlated with number of categories allocated to video.

To further validate the impact of colocation, Figure 14 presents the average number of views against the number of categories a video belongs to. We observe an almost linear trend in which videos belonging to multiple categories get more views. On average, videos without any category receive just 51k views, compared to 452k for those with 5 categories. We conclude that category tags appear to be a significant factor that contributes to views. It is quite possible that uploaders, particularly commercial ones, might exploit this observation. In fact, not being aware of the importance of categories can have dire consequences for video popularity. We observe that uncategorised videos appear in the bottom 20 most unpopular categories ranked by absolute viewing figures (and the least popular measured by normalised popularity).

6.4 Summary

We conclude that users in YouPorn rely heavily on category information for discovering content of interest. This appears to be primarily driven by category-based browsing; this observation is best highlighted by the almost linear relationship between the number of views a video receives and the number of categories it is

listed in. Obviously, this raises questions about exactly how inherently important categories are, as the tagging of a video in as many as 10 categories suggests a low level of accuracy in many cases. Instead, it seems probable that many users use categories as a very coarse way of targeting interest groups. Once again, this suggests a level of flexibility that is far less prevalent in traditional media types. We also find that this information is not being exploited well in YouPorn, with a poor level of market efficiency being shown in the corpus (i.e., often uploaders do not target their content well).

7. ELASTICITY IN CONTENT CONSUMPTION

Our results have highlighted a number of characteristics that might be collectively taken as implying that users do not visit YouPorn with specific videos in mind: duration does not matter (Figure 1); users tend to simply go for the most easily accessible videos (Figure 7); the number of views depends largely on the video being listed on the front page (Figure 9); and the number of views is correlated with the number of categories a video is listed in (Figure 14). These observations lead us to hypothesise that a significant portion of users are therefore quite flexible in what they watch.

To confirm this apparent elasticity in users' content consumption requirements (and to verify other findings), we performed a small-scale user study, recruiting 46 respondents over social networks and mailing lists. Consistent with our expectations, we find that 85% of users find it easy or just slightly difficult to find content of interest, with 15% saying they found it difficult. 43% of survey respondents also said that over $3/5$ of the videos they found match their interests.

We believe that this type of flexibility sheds light on some of the results obtained in the previous sections. Confirming our earlier suspicions, the survey shows that the observed findings do, indeed, seem to arise from users' dependencies on the order in which content is displayed on the default front page, and on category-specific pages: when asked what characteristics they use to find content (multiple answers were allowed), we found that only 22% of users ever visit the site with a pre-determined video in mind. Instead, a large share of the respondents appear to utilise what we have previously termed "flexible" ways to discover content: 63% rely on browsing the front page, 59% use category-based browsing, and 50% utilise the search functionality. Further, in-line with our earlier conjecture, we found that only 9% visit YouPorn through links from other sites. This is unlike the behavior of YouTube users, for instance, who rely more on web search engines [19] and external links [38], rather than browsing. It is also unlike other types of repositories, where users primarily visit to watch specific videos (e.g., sports games [11]).

When combining the above findings, it becomes likely that the more flexible users could all be satisfied with a relatively small set of videos taken from a large range of acceptable ones. We argue that this flexibility should therefore be leveraged by the content distribution infrastructure. Specifically, where many videos could satisfy the user, we posit that the content distribution infrastructure should guide users towards those particular ones that also have a low network cost (e.g., available nearby). This could improve user quality of experience and reduce network overhead, which benefits other users as well as ISPs. Considering users' predisposition towards browsing the front page, this could be done easily: instead of ordering videos by recency of upload (a common design pattern), the different browsing pages generated by the web front-end could also take into account which videos are available near to the viewer. Most simply, nearby videos that match a user's interests could be placed at the top of the front page.

Our user survey indicates that only a small amount of shared content would be required: 87% of users watch under 10 videos a session, while 43% watch 3 or less. Assuming a large intersection between these flexible users' interests, a relatively small amount of (generic) content would likely satisfy the demands of many viewers. This suggests the approach would be highly feasible in this domain. However, it must be ensured that optimising the delivery infrastructure is not done at the expense of a user's quality of experience. A key challenge here would be to ensure that this small amount of content is kept sufficiently "fresh". Only 24% of our survey respondents stated that they do not get bored easily and would watch a video multiple times. Thus, it is important that users are given sufficient choice to ensure that they are provided with novel items they would wish to watch.

8. SUMMARY AND FUTURE WORK

This paper presented the first detailed measurement study of Internet adult media distribution, focussing on the YouPorn website. Three key aspects of this system have been inspected: the corpus, the nature of content popularity and the impact of categories. We found that YouPorn is a hugely popular service with over 60 billion views recorded from a corpus of 183k videos. Unlike traditional UGC websites, there is an extremely prominent commercial element to its content, as well as a seemingly well managed vetting procedure. We observed a number of other interesting properties, particularly relating to the rapid decay of content popularity as measured from the number of views, as well as users' dependency on category metadata to find content that matches their interests. Further investigation uncovered the main reason behind these observations: the predominant use of YouPorn's browsing options. Particularly, this is driven by the apparent flexibility that most users have when accessing adult media: they do not seek a specific video, rather, they search for *any* video that falls with certain (broad) interest constraints. We posit that this is a characteristic that likely exists more generally in other multimedia repositories, but to varying extents. We therefore propose to exploit this observation by shaping users' browsing behaviour towards videos with a low network cost. In fact, this could be done with the intention of optimising any metric. The constantly expanding size and popularity of these repositories means that this is an approach that may become increasingly necessary to scale content delivery.

Due to its infancy, there is a significant amount of future work that could focus on adult video streaming. The dataset presented in this paper has focussed on aggregated system-level and video-level information. Whereas this offers insights into various corpus and popularity aspects, it does not provide user-level analysis. The next stage of our work will therefore focus on this type of data to understand exactly how individual users interact with such websites. This will capture their behaviour (e.g., skipping), as well as things like regional differences between user groups. Although not explored in this paper, our survey indicates that these elements are potentially quite different from traditional media. Such data will also allow us to gain a better understanding of things like traffic volumes and cacheability. We also intend to further develop the ideas explored in Section 7. Through user-level data, for example, we will be able to understand the current intersection of user interests and requests. More extensive user testing will also complement this. Beyond these targeted avenues of study, there are more general topics of interest, including social networking aspects and deeper analysis of the "2.0" elements of these services (e.g., ratings, comments, and recommendations).

9. REFERENCES

[1] Alexa. http://www.alexa.com/.

[2] Youporn. http://www.youporn.com.

[3] Youporn.com is now a 100% redis site. https://groups.google.com/forum/?fromgroups=#!topic/redis-db/d4QcWV0p-YM.

[4] Youtube statistics. http://www.youtube.com/yt/press/en-GB/statistics.html.

[5] Cisco visual networking index: Forecast and methodology, 2011-2016, 2012.

[6] ABRAHAMSSON, H., AND NORDMARK, M. Program popularity and viewer behaviour in a large TV-on-demand system. In *Proc. ACM IMC* (2012).

[7] ANTHONY, S. Just how big are porn sites? http://www.extremetech.com/computing/123929-just-how-big-are-porn-sites.

[8] APOSTOLOPOULOS, J. G., TAN, W.-T., AND WEE, S. J. Video streaming: Concepts, algorithms, and systems. *HP Laboratories, report HPL-2002-260* (2002).

[9] ATTWOOD, F. *Porn.com: Making sense of online pornography*, vol. 48. Peter Lang, 2010.

[10] BORGHOL, Y., ARDON, S., CARLSSON, N., EAGER, D., AND MAHANTI, A. The untold story of the clones: Content-agnostic factors that impact YouTube video popularity. In *Proc. of ACM SIGKDD* (2012).

[11] BRAMPTON, A., MACQUIRE, A., FRY, M., RAI, I. A., RACE, N. J., AND MATHY, L. Characterising and exploiting workloads of highly interactive video-on-demand. *Multimedia systems 15*, 1 (2009), 3–17.

[12] CARROLL, J. S., PADILLA-WALKER, L. M., NELSON, L. J., OLSON, C. D., BARRY, C. M., AND MADSEN, S. D. Generation XXX: Pornography acceptance and use among emerging adults. *Journal of adolescent research 23*, 1 (2008), 6–30.

[13] CHA, M., KWAK, H., RODRIGUEZ, P., AHN, Y.-Y., AND MOON, S. Analyzing the Video Popularity Characteristics of Large-scale User Generated Content Systems. *IEEE/ACM Trans. Netw. 17*, 5 (2009), 1357–1370.

[14] CHA, M., RODRIGUEZ, P., CROWCROFT, J., MOON, S., AND AMATRIAIN, X. Watching television over an IP network. In *Proc. ACM IMC* (2008).

[15] CHENG, X., DALE, C., AND LIU, J. Statistics and social network of YouTube videos. In *Proc. IEEE IWQoS* (2008).

[16] COOPER, A. Sexuality and the internet: Surfing into the new millennium. *CyberPsychology & Behavior 1*, 2 (1998), 187–193.

[17] COOPERSMITH, J. Does your mother know what you really do? The changing nature and image of computer-based pornography. *History and Technology 22*, 1 (2006), 1–25.

[18] CRANE, R., AND SORNETTE, D. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of The National Academy of Sciences 105*, 41 (2008), 15649–15653.

[19] CUNNINGHAM, S. J., AND NICHOLS, D. M. How people find videos. In *Proc. ACM/IEEE JCDL* (2008).

[20] DANEBACK, K., SEVCIKOVA, A., MÅNSSON, S.-A., AND ROSS, M. W. Outcomes of using the internet for sexual purposes: Fulfilment of sexual desires. *Sexual Health 10* (2012), 26–31.

[21] DINES, G. *Pornland: How Porn Has Hijacked Our Sexuality*. Beacon Press, 2010.

[22] GAO, P., LIU, T., CHEN, Y., WU, X., ELKHATIB, Y., AND EDWARDS, C. The measurement and modeling of a P2P streaming video service. In *GridNets* (2008).

[23] GUO, L., TAN, E., CHEN, S., XIAO, Z., AND ZHANG, X. The stretched exponential distribution of internet media access patterns. In *Proc. ACM PODC* (2008).

[24] HEI, X., LIANG, C., LIANG, J., LIU, Y., AND ROSS, K. W. A measurement study of a large-scale P2P IPTV system. *IEEE Trans. Multimedia 9*, 8 (2007), 1672–1687.

[25] HU, W., WU, O., CHEN, Z., FU, Z., AND MAYBANK, S. Recognition of pornographic web pages by classifying texts and images. *IEEE Trans. Pattern Analysis and Machine Intelligence 29*, 6 (2007), 1019–1034.

[26] HURLEY, R., PRUSTY, S., SOROUSH, H., WALLS, R. J., ALBRECHT, J., CECCHET, E., LEVINE, B. N., LIBERATORE, M., LYNN, B., AND WOLAK, J. Measurement and analysis of child pornography trafficking on P2P networks. In *Proc. WWW* (2013).

[27] LABOVITZ, C., LEKEL-JOHNSON, S., MCPHERSON, D., OBERHEIDE, J., AND JAHANIAN, F. Internet inter-domain traffic. In *Proc. SIGCOMM* (2010).

[28] MEHTA, M. D., AND PLAZA, D. Content analysis of pornographic images available on the internet. *The Information Society 13*, 2 (1997), 153–161.

[29] NENCIONI, G., SASTRY, N., CHANDARIA, J., AND CROWCROFT, J. Understanding and decreasing the network footprint of over-the-top on-demand delivery of TV content. In *Proc. WWW* (2013).

[30] OGAS, O., AND GADDAM, S. *A billion wicked thoughts: what the world's largest experiment reveals about human desire*. Dutton, 2011.

[31] PAASONEN, S. *Carnal Resonance: Affect and Online Pornography*. MIT Press, 2011.

[32] PINTO, H., ALMEIDA, J. M., AND GONÇALVES, M. A. Using early view patterns to predict the popularity of YouTube videos. In *Proc. ACM WSDM* (2013).

[33] SASTRY, N. How to tell head from tail in user-generated content corpora. In *Proc. ICWSM* (2012).

[34] SCHUHMACHER, M., ZIRN, C., AND VÖLKER, J. Exploring YouPorn categories, tags, and nicknames for pleasant recommendations. In *Proc. SEXI* (2013).

[35] SCHULZE, H., AND MOCHALSKI, K. Internet study 2007-2009. ipoque Report, 2009.

[36] SULER, J. The online disinhibition effect. *Cyberpsychology & Behavior 7*, 3 (2004), 321–326.

[37] SZABO, G., AND HUBERMAN, B. A. Predicting the popularity of online content. *CACM 53*, 8 (2010), 80–88.

[38] WATTENHOFER, M., INTERIAN, Y., VAVER, J., AND BROXTON, T. Catching a viral video. In *Proc. Workshop on Social Interaction Analysis and Service Providers* (2010).

[39] WONDRACEK, G., HOLZ, T., PLATZER, C., KIRDA, E., AND KRUEGEL, C. Is the internet for porn? An insight into the online adult industry. In *Proc. Workshop on Economics of Information Security* (2010).

[40] YU, H., ZHENG, D., ZHAO, B. Y., AND ZHENG, W. Understanding user behavior in large-scale video-on-demand systems. *ACM SIGOPS Operating Systems Review 40*, 4 (2006), 333–344.

[41] ZINK, M., SUH, K., GU, Y., AND KUROSE, J. Characteristics of youtube network traffic at a campus network–measurements, models, and implications. *Computer Networks 53*, 4 (2009), 501–514.