Project Description

Kenneth Morales

October 5, 2017

*IMPORTANT NOTE: ALL CAPS AND BOLDED TEXT ONLY PERTAINS TO THIS PRELIMINARY ANALYSIS.* *THIS ANALYSIS WAS PERFORMED ON MY ORIGINAL "TEST" SCRAPING DATASET, WHICH WAS FROM THE SITE INDEX BY CATEGORY. IT WAS LAID OUT IN AND EASY TO READ AND SCRAPE TABLE, SO I TOOK IT AND THEN NAVIGATE TO EACH VIDEO TO GET ADDITIONAL DETAILS* *I AM STILL HAVING TROUBLE GETTING A "RANDOM" DATASET BY PORNHUB'S "RANDOM VIDEO" FEATURE. PLEASE SEE THE SECTION ON SCRAPING THE DATA AND ADVISE IF ABLE.*

### Question

Is there a correlation between the safer sex practices pertaining to condoms employed in gay porn videos viewed on the popular pornographic website PornHub (PH) in the United States in the last 5 - 10 years and the new guidelines recommended for Truvada to be used as pre-exposure prophylaxis (PrEP) on May 14, 2014?

Part I. A (I think it is exploratory?) data analysis of gay male pornographic videos available online and their viewing trends over the past 5-10 years. *PART I.5. A REGRESSION-DISCONTINUITY ANALYSIS ON THE RELATIVE POPULARITY OF "BAREBACK" CATEGORY VIDEOS OVER TIME* Part II. A Regression-Discontinuity analysis on the relative popularity of "Bareback"" category videos over time: *THIS HAS BEEN SET ASIDE UNTIL I GET THE FINAL DATASET*

Variable of interst: Films designated as "bareback" (no visible condom use) = 0 for no, 1 for yes. Classified as such by: *FOR NOW, I AM JUST STICKING WITH THE CATEGORY LABEL PROVIDED BY PORNHUB* Classified as such by: identifying a string pattern of gay slang terms commonly used to describe penetrative sex without condoms Measures: Ratio of mean views of films by bareback category "Program" group: Films added to PH after the cutoff, as chosen by a random sample (N = 100000) "Control" group: Films added to PH before the cutoff Cutoff: May 14, 2014 (publication of guidelines for recommendation of Truvada as PrEP in the USA) + an unknown amount of lag time?

Source: Stephen pointed me to the [Regression-Discontinuity design](https://www.socialresearchmethods.net/kb/quasird.php).

Throughout this analysis, "bareback" will refer to sex between two men without the use of condoms, as it is the most common term for such an activity in gay verbiage.

### Background:

#### PornHub

[PornHub](www.pornhub.com) is the world's largest pornography site on the internet, operating now for 10 years. Currently, it is the 39th most popular site on the web. Operating similarly to YouTube, PornHub claims itself as a "platform" and a "video host:" in such a way, PornHub is capable of reaping the benefits of the massive traffic generated to its site while claiming no responsibility for the videos therein. In 2010, the start-up was bought out by a large adult entertainment conglomerate Manwin (now known as MindGeek), who owns several other similar websites.

Started in 2013, [PornHub Insights](https://www.pornhub.com/insights) is a blogging platform wherein the massive quantities of data reaped from PornHub users would be periodically analyzed. Insights posts were generally themed around a particular large event, monitoring traffic changes around NBA finals, episodes of Game of Thrones, or the path of totality for the 2017 solar eclipse. Occassionally, and at the end of every year, PronHub Insights will break down data more generally, by geography, times of access, popularity of search terms and categories, etc.

#### PrEP

Pre-exposure prophylaxis, or PrEP, is the preemptive use of a drug to prevent disease in unexposed populations. Currently, the term is used almost exclusively as shorthand for the use of the antiretroval drug Truvada, a two-drug combo manufactured by Gilead Sciences, to prevent the replicaton of HIV. Truvada's regimen as of the time of writing is a single pill taken daily coupled with a follow-up visit with the patient's primary care provider every three months, presumably in perpetuity.

Truvada has had the most adoption success among men who have sex with men in the United States after approval by the CDC for use as PrEP in 2014. Moves toward widespred adoption of PrEP have been divisive, politically and within gay culture. Concerns include the likelihood of PrEP being used counter to the prescription (such as in the phenomenon of "disco dosing," an ineffective practice in which men take a Truvada pill only around the time they expect sex), the potential for PrEP to undermine existing safer sex policies and social mores, and the incredible cost of the drug (as high as $450 / month). As of 2017, Truvada as PrEP has been approved by 9 countries in addition to the United States as well as the World Health Organization.

#### Gay Porn

Since the AIDS crisis in the 1980s, condoms have been standard parts of gay pornographic films. While condom-less gay porn ("barebacking") still did exist, it was commonly regarded as kink, fetish, or otherwise deviant, and was often the domain of specialized or heterodox smaller production studios.

In recent years, porn studios, both gay and straight, have seen their profits being eaten in to by a combination of three phenomenons:

1. The rise of the "tube" sites (such as PornHub), which aggregate and disseminate pornography, usually without a membership fee,
2. The piracy of pornography and distribution via peer-to-peer networks and forums, as well as through the tune sites, and
3. burgeoning competition from amateur, DIY pornographers, many of which are capable of providing bareback porn, being unrestrained by mainstream conventions.

With the advent of PrEP, however, gay porn studios suddenly had an out: they could provide bareback sex for audiences who preferred it while still bearing the mantle of "safer" sex. Most large gay pornographic studios began producing bareback porn in the years following the CDC's endorsement of Truvada as PrEP.

#### Obtaining the raw data

Performed on October 6th 2017 (code finished running on October 8th, 2017) using the "Random video" function of PH's website. I will only be analyzing videos that are not behind PH's paywall for subscribers to PH Premium (HD videos and an advertisement-free interface). Do not run this code: the dataset will be uploaded by following codeblocks. Unfortunately, I am unable to grab the video lengths, though that is not necessary for my analysis:

*PROFESSORS PLEASE ADVISE ON MY SCRAPE CODE. I AM STILL UNSUCCESSFUL AT GETTING DATA--UNSURE IF IT IS USER AGENT OR ISP-RELATED OR WHAT* *I AM POSTING SCRAPING BY RANDOM VIDEO CODE HERE. I ALSO HAVE THE ORIGINAL CODE I USED TO SCRAPE THE DATA USED FOR THIS ANALYSIS (FROM THE SITE INDEX BY CATEGORY). LET ME KNOW IF YOU WANT TO SEE THAT, BUT SINCE IT SOUNDS LIKE Y'ALL AREN'T ALL THAT INTERESTED IN THE DATA SCRAPING PORTION I AM JUST NOT GOING TO INCLUDE IT!*

# Randomly grab 10000 videos using PH's "random" link  
  
url <- "https://www.pornhub.com/gay/video/random"  
trimpatdate <- dgt(6) %R% "/" %R% dgt(2)  
trimpatvk <- fixed("https://www.pornhub.com/view\_video.php?viewkey=")  
randviddat.list = list()  
x <- 10000  
  
# Seem unable to extract length from video  
# length <- html\_nodes(webpage, '.mhp1138\_total') %>% html\_text()  
  
# Might need to mess with User session stuff!  
  
for(i in 1:x) {  
 Sys.sleep(runif(1,0,5))  
 session <- html\_session(url, user\_agent(uastring))  
 if (session[[6]][2] == 200) {  
 webpage <- read\_html(session[[3]])  
 if (length(html\_attr(html\_nodes(webpage, '.premiumLocked'), name = "class") != 0)) {  
 x <- x+1  
 } else {  
 title <- html\_nodes(webpage, '.inlineFree') %>% html\_text()  
 views <- html\_nodes(webpage, '.count') %>% html\_text() %>% str\_replace\_all("[,]", "") %>% as.numeric()  
 rating <- html\_nodes(webpage, '.percent') %>% html\_text() %>% str\_replace("[%]", "") %>% as.numeric()  
 categories <- html\_nodes(webpage, '.categoriesWrapper > a') %>% html\_text() %>% str\_c(collapse = ", ")  
 production <- html\_nodes(webpage, '.production') %>% html\_text()  
 tags <- html\_nodes(webpage, '.tagsWrapper > a') %>% html\_text() %>% str\_c(collapse = ", ")  
 added <- html\_nodes(webpage, xpath = '/html/head/link[5]') %>% html\_attr(name = "href") %>% str\_extract(pattern = trimpatdate) %>% str\_replace("/", "")  
 viewkey <- html\_nodes(webpage, xpath = '/html/head/link[4]') %>% html\_attr(name = "href") %>% str\_replace(pattern = trimpatvk, "")  
 randviddat.list[[i]] = data.frame(title,views,rating,categories,production,tags,added,viewkey)  
 }  
 } else {  
 x <- x+1  
 }  
}  
  
# Compile into single dataframe, identifies duplicates, saves data  
all\_vids = do.call(rbind,randviddat.list)  
duplicate <- duplicated(all\_vids[,7])  
all\_vids[,"duplicate"] <- duplicate  
save(all\_vids, file = "data/randvid\_data\_dup.Rda")  
all\_vids <- subset(all\_vids, !duplicated(all\_vids[,7]))  
all\_vids <- all\_vids[,-ncol(all\_vids)]  
save(all\_vids, file = "data/randvid\_data.Rda")

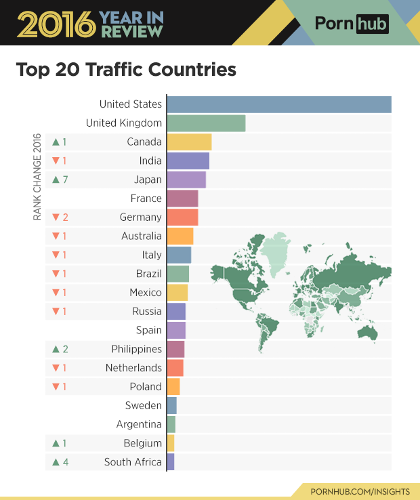
###### Limitations:

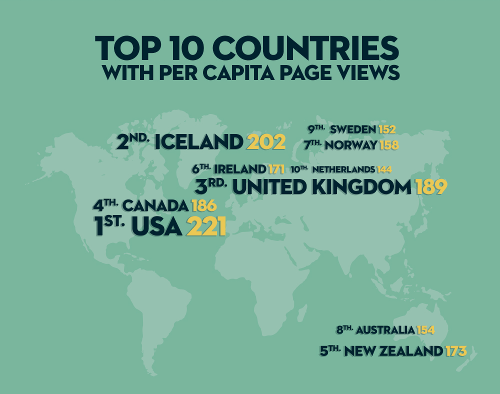
1. Data by Country Viewership

Unfortunately, due to [PornHub's robots.txt](https://www.pornhub.com/robots.txt), I am unable to specifically scrape by country code:

Disallow: /\*cc=\*

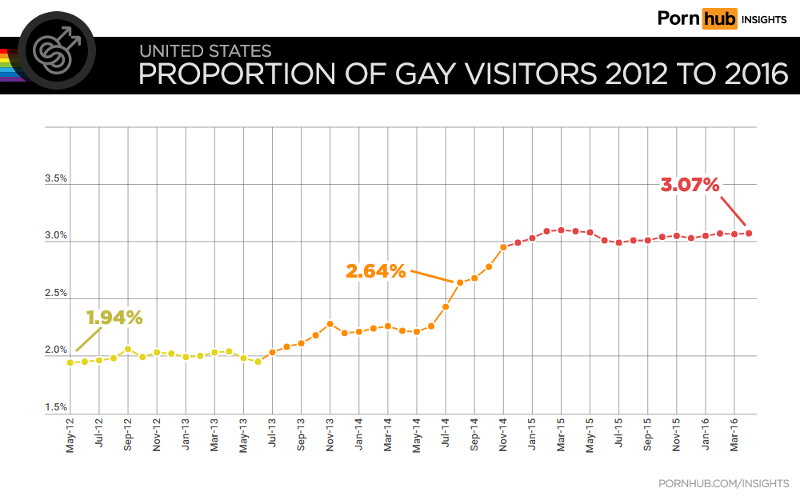
However, their own published web traffic data by country allows for a reasonable assumption that the broader trends found without incorporating country code could reflect USA-specific viewing trends:





1. Viewership of videos by MSM

PornHub provides a specific [url stem](https://www.pornhub.com/gayporn) to cordon off "gay" porn from the main site's "straight" porn, the top-level domain. Essentially, the category "Gay" is applied to videos in this section of the website, as a high-level video filter. For the purpose of this analysis, the assumption is made that MSM would seek out porn from this gay porn section of the website. According to PornHub's own usage statistics for 2016, ~3% of visitors from the US are "gay visitors," a statistic that is commensurate with national surveys on American's sexual orientation.



1. The Paywall Effect

For my analysis, I was unable to scrape data from videos behind PH's Premium service paywall. It is possible that viewing habits and the videos themselves might differ between Premium subscribers and users who only access content not behind a paywall. However, according to CovenantEyes, a religiousky-based anti-porn advocacy organization modeled after consumer watchdogs, 9 out of 10 users access free pornography.

1. Reliance on text descriptors of videos

For this analysis, I am relying on text descriptors, both user-supplied (video title, tags) and host-supplied (categories), in order to determine if a video contains sex between two or more performers without use of condoms. Addtionally, for the "Solo Male" category, which is removed from the sample due to its inability to capture the activity of interest (penetrative sex), there is a possibility for misclassification also. This introduces a potential for incorrect classification, but viewing individual videos is outside of the scope of this project.

1. Reliance on "random" video URL *DOES NOT PERTAIN TO THE ANALYSIS HERE, WHICH WAS DONE BY THE CATEGORICAL SITE INDEX (500 MOST RECENTLY "FEATURED" VIDEOS, HOWEVER THAT IS DECIDED*

I have no way of knowing how PH decides which video will appear on its "random" video service, but I am assuming it is truly random for purposes of this project.

*6. UPLOAD DATE BINNED BY YEAR* *PORNHUB DOES A REALLY ANNOYING THING WHERE THEY STORE UPLOAD DATES HARD CODED AS SOMETHING LIKE: "ADDED 2 YEARS AGO." I DIDN'T FIGURE OUT HOW TO GRAB THE ACTUAL UPLOAD DATES UNTIL AFTER I HAD SCRAPED THE INITIAL DATASET UPON WHICH THIS ANALYSIS IS BASED. SO ALL OF THESE ANALYSES ARE DONE ON A YEAR INTEGER VARIABLE, NOT ANY OTHER BINNED TIME UNIT.*

#### Data Cleaning & Processing

##### Exploring Pornhub's Categories

Stuff about cleaning the dates up.

# Need to insert data code once I get access to the randomized datset with true upload dates

Load in the data. *FIX THIS WITH REAL DATA SET DO NOT FORGET!!!*

load("~/adsfinalproj/data/siteindexvidyear\_data.Rda")  
all\_vids <- bysiteindex\_data  
# Capture n for sample table  
n1 <- nrow(all\_vids)

Videos on PH have a system-defined organization called "Categories." A quick word cloud will give us a sense of what the category breakdown is for the videos that have been scraped:

# Wordcloud of Categories  
cats <- str\_split(all\_vids$categories, ", ")  
cats <- data.frame(category = names(table(unlist(cats))), freq=as.numeric(table(unlist(cats)))) %>% arrange(desc(freq))  
cats$category <- as.character(cats$category)  
wordcloud(words = cats$category, freq = cats$freq, scale = c(6,1), random.order=FALSE, rot.per=0.1, colors=brewer.pal(8, "Dark2"))

## Warning in wordcloud(words = cats$category, freq = cats$freq, scale =  
## c(6, : Virtual Reality could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = cats$category, freq = cats$freq, scale =  
## c(6, : Verified Models could not be fit on page. It will not be plotted.



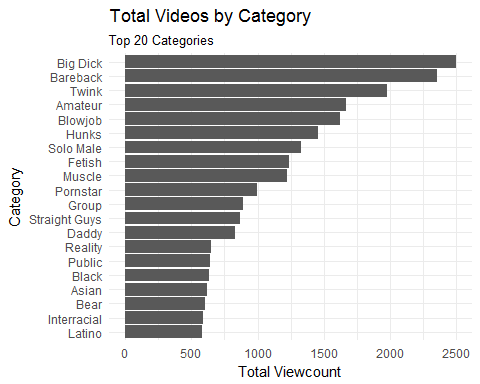
Several of these categories would be useless to analyze: some specify the video format ("HD," "Virtual Reality"), the type of camerawork ("POV," "Webcam"), or the performer's relationship to PH itself ("Verified Models," "Verified Amateurs," "Exclusive"). Additionlly, the "Gay" category tag is useless for this analysis, since that is how the PH website filters videos for the straight and gay porn domains.

# Remove analytically useless categories  
badcats <- "Gay|HD|Virtual Reality|Verified Amateurs|Verified Models|Exclusive|POV|Webcam"  
cats <- cats[!str\_detect(cats$category, badcats), ]  
wordcloud(words = cats$category, freq = cats$freq, scale = c(2.5,1), random.order=FALSE, rot.per=0.1, colors=brewer.pal(8, "Dark2"))



The category "Bareback" appears to be one of the most popular video categories:

ggplot(data = cats, aes(x = category, y = freq)) + theme\_minimal() + scale\_color\_brewer(palette="Paired") + geom\_col() + coord\_flip() + scale\_x\_discrete(limits=rev(cats$category[1:20])) + theme(legend.position = "bottom") + labs(title = "Total Videos by Category", subtitle = "Top 20 Categories", x="Category", y = "Total Viewcount")



Additionally, one category sticks out as being contrary to the analysis: "Solo Male." Videos with only one performer are incapable of containing either condom-utlizing or condomless sex.

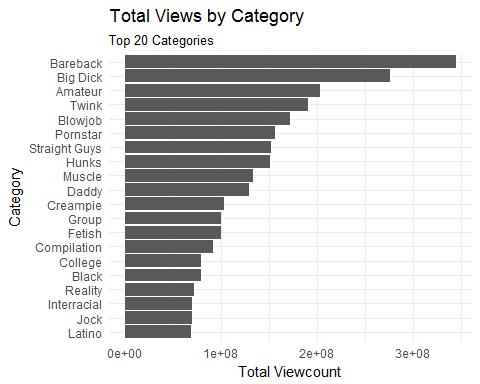
# Remove videos with only a single male performer in them as identified by category  
nosolos <- all\_vids[- grep("Solo Male", all\_vids$categories),]  
# Capture n for sample table  
n2 <- nrow(nosolos)

##### Popularity Rankings as Given by View Count by Category

Popularity ranking reveals categories by the number of views generated by all videos in a given category over a time span, weighted by the number of these videos. This shows the repetition of views on videos in a given category, revealing the consistency of viewers’ requests for this content. These rankings tell us what categories were the most popular, and may point to content for which demand surpasses what is offered by uploaders.

To begin with, we will identify the most popular categories as given by view count of videos over the entire sample:

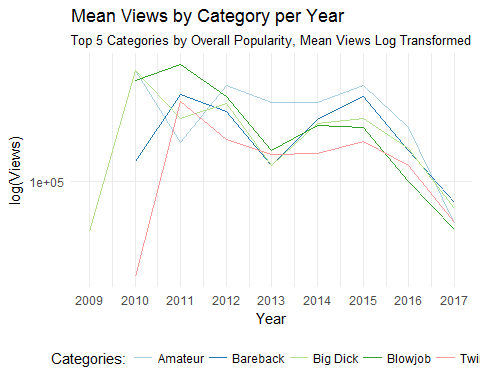
# Calculate mean views by category and year  
poprank <- cats[,1]  
for(i in seq\_along(poprank)) {  
 if(i == 1){  
 poprankyr <- data.frame(nosolos %>% group\_by(year) %>% subset(subset = str\_detect(categories, poprank[i])) %>% summarise(cat = poprank[i], n = n(), mviews = mean(views)))  
 }  
 if(i >= 2) {  
 poprankyr <- rbind(poprankyr, data.frame(nosolos %>% group\_by(year) %>% subset(subset = str\_detect(categories, poprank[i])) %>% summarise(cat = poprank[i], n = n(), mviews = mean(views))))  
 }  
}  
  
# Identify the most popular categories overall  
topcats <- poprankyr %>% group\_by(cat) %>% summarise(tviews = sum(mviews\*n)) %>% arrange(desc(tviews))  
  
ggplot(data = topcats, aes(x = cat, y = tviews)) + theme\_minimal() + scale\_color\_brewer(palette="Paired") + geom\_col() + coord\_flip() + scale\_x\_discrete(limits=rev(topcats$cat[1:20])) + theme(legend.position = "bottom") + labs(title = "Total Views by Category", subtitle = "Top 20 Categories", x="Category", y = "Total Viewcount")



It appears that while Bareback was one of the most popular categories for video uploads, it is the most viewed category overall.

A quick glimpse at the rankings over time of the top 5 most viewed overall categories:

topcats <- as.data.frame(topcats)  
popcats <- topcats[1:5,1]  
popcatspat <- str\_c(popcats, collapse = "|")  
popranktop5 <- poprankyr[grep(popcatspat, poprankyr$cat),]  
  
# By year plot  
ggplot(data = popranktop5, aes(x = year, y = mviews)) + theme\_minimal() + scale\_color\_brewer(palette="Paired") + geom\_line(aes(color = cat)) + scale\_x\_continuous(breaks=seq(min(popranktop5$year),max(popranktop5$year),1)) + scale\_y\_log10() + theme(legend.position = "bottom") + labs(title = "Mean Views by Category per Year", subtitle = "Top 5 Categories by Overall Popularity, Mean Views Log Transformed", x="Year", y = "log(Views)", color = "Categories:")



##### Constructing Key Variable: Presence of Penetrative Sex without a Condom

*WHILE I STILL CONSTRUCT AND ANALYZE THIS VARIABLE, IT WILL NOT PLAY IN THE FINAL ANALYSIS*

Thus far only the PH-dictated organizational method has been explored (category). User-submitted categorization methods (titles of videos and "tags") will be scanned for terms commonly associated with sex without condoms in gay slang. Though "Bareback" is a category of videos, this category may not fully capture all videos that contain penetrative sex without the use of condoms.

# Generate a pattern to detect for sex without condoms  
bb\_pattern <- "bare|(?:^| )bb(?:$| )|breed|cream|cum dump|felch|raw|condomless|no condom"  
  
# Scan titles, tags, and categories for indications of barebacking  
nosolos$categories <- as.character(nosolos$categories)  
nosolos$titles <- as.character(nosolos$titles)  
nosolos$tags <- as.character(nosolos$tags)  
  
bb\_vids\_titles <- nosolos[grepl(bb\_pattern, nosolos$titles, ignore.case=TRUE),]  
bb\_vids\_cats <- nosolos[grepl(bb\_pattern, nosolos$categories, ignore.case=TRUE),]  
bb\_vids\_tags <- nosolos[grepl(bb\_pattern, nosolos$tags, ignore.case=TRUE),]  
  
# Generate a binary variable "bareback" for videos containing indicatons of sex without condoms  
bbvids <- rbind(bb\_vids\_cats, bb\_vids\_tags, bb\_vids\_titles, stringsAsFactors = FALSE)  
bbvids <- unique(bbvids)  
bbvids$bareback <- 1  
  
notbbvids <- anti\_join(nosolos, bbvids)

## Joining, by = c("viewkey", "category", "titles", "vid\_length", "views", "rating", "url", "categories", "production", "tags", "added", "dateadded", "year")

notbbvids$bareback <- 0  
nosolos <- rbind(bbvids, notbbvids, stringsAsFactors = FALSE)  
  
# Make binary bareback variable in data factors  
nosolos$bareback <- factor(nosolos$bareback, labels = c("No", "Yes"))  
  
catsns <- nosolos[c("categories", "year", "views", "bareback")]  
catsns$bbcat <- factor(ifelse(str\_detect(catsns$categories, "Bareback"), 1, 0), labels = c("No", "Yes"))  
bbcomp <- data.frame(table(catsns$bbcat, catsns$bareback))  
names(bbcomp) <- c("BB Category", "BB Pattern", "Freq")  
# Need to pretty this up  
kable(bbcomp, format = "html", booktabs = T)

BB Category

BB Pattern

Freq

No

No

5791

Yes

No

0

No

Yes

1208

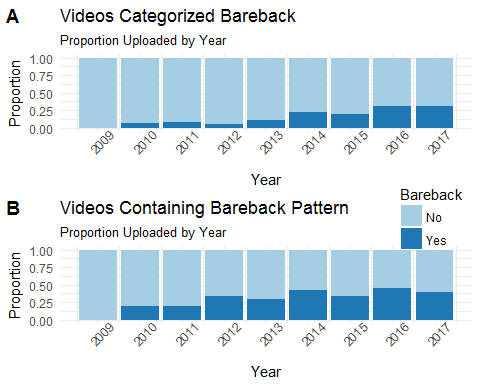
Yes

Yes

2354

There is substantial divergence from the Bareback categorized videos and those identified by the pattern.

# Compare Number of Videos Categorized Bareback to those Identified as BB  
# By count  
catbbcnt.plot <- ggplot(data = catsns, aes(x = year, fill = bbcat)) + theme\_minimal() + geom\_bar(position = "fill") + scale\_fill\_brewer(palette="Paired") + scale\_x\_continuous(breaks=seq(min(catsns$year),max(catsns$year),1)) + theme(axis.text.x = element\_text(angle=45)) + theme(legend.position="none") + labs(title = "Videos Categorized Bareback", subtitle = "Proportion Uploaded by Year", x="Year", y = "Proportion", fill = "Bareback")  
isbbcnt.plot <- ggplot(data = catsns, aes(x = year, fill = bareback)) + theme\_minimal() + geom\_bar(position = "fill") + scale\_fill\_brewer(palette="Paired") + scale\_x\_continuous(breaks=seq(min(catsns$year),max(catsns$year),1)) + theme(axis.text.x = element\_text(angle=45)) + theme(legend.position= c(0.9,1.35)) + labs(title = "Videos Containing Bareback Pattern", subtitle = "Proportion Uploaded by Year", x="Year", y = "Proportion", fill = "Bareback")  
  
plot\_grid(catbbcnt.plot, isbbcnt.plot, labels = c("A", "B"), nrow = 2, align = "v", rel\_heights = c(1,1))



*AS YOU CAN SEE HERE, THERE IS A QUITE A DIFFERENCE BETWEEN THE PROPORTIONS OF VIDEOS UPLOADED BY YEAR BETWEEN PORNHUB'S BAREBACK CATEGORY AND MY OWN BAREBACK PATTERN RECOGNITION SYSTEM, WHICH IS INTERESTING WHEN YOU CONSIDER, FOR EXAMPLE, THAT NOT ONLY A) IF MY HYPOTHESIS WAS CORRECT, USERS MIGHT HAVE SOUGHT OUT / UPLOADED MORE BAREBACK CATEGORY VIDEOS AFTER THE CDC GUIDELINES POSTED, AND B) THAT THE "GAY" CATEGORY IS HOW PORNHUB FILTERS THE GAY PORN (WWW.PORNHUB.COM/GAYPORN) FROM THE "STRAIGHT" PORN (WWW.PORNHUB.COM)*

The dataset will undergo a final polish pass before analysis.

cleandata <- nosolos[c("titles", "categories", "tags", "year", "views", "rating", "production", "viewkey", "url", "bareback")]  
cleandata <- arrange(cleandata, year)

*I WILL REINTRODUCE THE DUMMY VARIABLE FOR PH'S BAREBACK CATEGORY HERE, TO POTENTIALLY REMOVE IN FINAL ANALYSIS, OR TO ALSO ANALYZE BY REGRESSION DISCONTINUITY*

cleandata$bbcat <- factor(ifelse(str\_detect(cleandata$categories, "Bareback"), 1, 0), labels = c("No", "Yes"))

#### Analysis - Regression Discontinuity

##### Specifications

Y = the ratio of means of the total views for videos in a given year, by bareback signifier where the bareback signifier is defined as: *THE CATEGORY "BAREBACK" AS SCRAPED FROM PH*

meanviews <- cleandata %>% group\_by(year, bbcat) %>% summarise(mviews = mean(views))

X = *YEAR VIDEO FILE WAS UPLOADED* Z = dummy variable for cutoff grouping (0 = pre-cutoff group, 1 = post-cutoff group)

1. The Cutoff Criterion: For the purposes of this analysis, the publication of the CDC guidelines recommending Truvada for PrEP among high-risk groups of individuals (serodiscordant couples, MSM who have sex without a condom or has been diagnosed with a sexually transmitted infection in the past 6 months and who is not in a mutually monogamous relationship with a partner who has recently tested negative for HIV, heterosexual people "at high risk" for contracting HIV from partner, and people who inject drugs who have shared equipment when injecting in the last 6 months): May 14, 2014. *NOTE: possibility I may incorporate some sort of lag time after the announcement, but how to decide on that span of time is beyond me at the moment.*

# IMPORTANT: For the rough draft, the cutoff is January 1, 2014, as the variables are examined in year bins--full dates will be available if I am ever successful at scraping the other dataset  
# Dropping the only video posted in 2009  
cleandata <- cleandata[-1,]  
# Set cutoff to 2014   
cutoffscore <- 2014  
cleandata$cutoff <- ifelse(cleandata$year < cutoffscore, 0, 1)

1. The Pre-Post Distribution It is assumed that the pre-post distribution is describable as a polynomial function.
2. Comparison Group Pretest Variance There must be a sufficient number of pretest values in the comparison group to enable adequate estimation of the true relationship (i.e., pre-post regression line) for that group. While the groups will be uneven, the sample size of videos posted before the cutoff is sufficient.

cleandata %>% group\_by(cutoff) %>% summarise(n = n()) %>% ungroup ()

## # A tibble: 2 x 2  
## cutoff n  
## <dbl> <int>  
## 1 0 1427  
## 2 1 7925

1. Continuous Pretest Distribution Both groups must come from a single continuous pretest distribution with the division between groups determined by the cutoff. Videos from PH will not differ "naturally" at the cutoff prior to the "program."
2. Program Implementation In this instance, since the implementation of the "program" is simply the passing of a date threshold, it is assumed that the program is uniformly delivered to all videos.

##### Analysis Strategy

Will overspecify the model in order to yield an unbiased estimate of the program effect, even though it may be ineffiect if the model indeed overspecifies. Will remove higher-order terms until the program effect estimate appears to different from the initial one or until residual plots indicate that the model fits poorly.

##### Steps in Analysis

1. Transform the Pretest The analysis begins by subtracting the cutoff value from each pretest score, creating the modified pretest term shown in Figure 7. This is done in order to set the intercept equal to the cutoff value (i.e., set X to 0 at the cutoff, making the cutoff the intercept point).

cleandata$pretest <- cleandata$year - cutoffscore

1. Examine Relationship Visually 2a. Determine whether there is any visually discernable discontinuity in the relationship at the cutoff: change in level vertically (main effect); change in slope (interaction effect); both.

2b. Degree of polynomial that may be required, as indicated by bivariate slope of the distribution.

1. Specify Higher-Order Terms and Interactions. By flexion points detected in step 2, creates transformations of the modified assignment variable. X. Transformation: two orders of polynomial higher than was indicated by the number of flexion points.
2. Estimate Initial Model Depending on the number of flexion points detected in step 2, one next creates transformations of the modified assignment variable, X. The rule of thumb here is that you go two orders of polynomial higher than was indicated by the number of flexion points. Thus, if the bivariate relationship appeared linear (i.e., there were no flexion points), one would want to create transformations up to a second-order (0 + 2) polynomial. This is shown in Figure 8. There do not appear to be any inflexion points or "bends" in the bivariate distribution of Figure 8.

#### Results

*SHOOT HIT THE FAN, LOOK FORWARD TO IT NEXT WEEK*

#### Discussion

TBA

#### References

Need to format!

<https://www.cdc.gov/nchhstp/newsroom/2014/PrEP-Guidelines-Press-Release.html>

END