Consumption of Pornography Online[*]

Lucas Shen[†]   Gaurav Sood[‡]

December 10, 2022

NB: Preliminary draft. Please do not cite without permission.

Abstract

Consumption of pornography has been blamed for a variety of societal ills, including, the rise in misogyny, sex crimes, and coarsening of the culture. Using passive browsing data from YouGov, we investigate how much pornography Americans consume online. We find that there is a sharp positive skew in the consumption of pornography, with a small number of users consuming lots of pornography with most consuming small amounts. The median American Internet user today spends X% of time-consuming pornography, visiting Y sites per month. Lastly, we find that, unlike previous research, relying on ecological inference, liberals consume slightly more pornography than conservatives (**??**).

Consumption of pornography is associated with a variety of disturbing attitudes, beliefs, emotions, and behaviors. Consuming pornography is associated with support for violence against women (**???**), belief in rape myths (**?**), greater gender role conflict and lesser sexual satisfaction (**??**), poorer relationship quality (**??**), and sexually risky behaviors such as engaging in paid sex, and having extramarital sex (**?**). A lot of popular pornography also contains a healthy dose of violence. One analysis of popular pornography revealed that 88.2% of the analyzed scenes contained physical aggression, and 48.7% verbal aggression (**?**). In all, there exist good reasons to be worried about rise in consumption of pornography.

In this paper, using passive browsing data from comScore and a YouGov sample, we investigate how much pornography do Americans consume online, how consumption of pornography online has changed over time, and shed light on who consumes pornography. We find that there is sharp skew in the consumption of pornography, with a small set of users consuming a large chunk of pornography. The median Internet user today spends X% of time consuming pornography, and visits Y sites per year. Between 2002 and 2014, the consumption of pornography online has increased by between XX and YY%. Lastly, we find that, unlike previous data, relying on ecological inference, liberals consume more pornography than conservatives (**??**).

(**?**) (**?**)

Data

We exploit data from comScore and a YouGov survey to measure the consumption of adult content.

comScore maintains a panel of around 100,000 households, offering small rewards in exchange for monitoring all Internet data from a machine. There are multiple notable biases in the sample. See **??** for comparison between demographics and population benchmarks on

some key variables.

YouGov maintains a large online panel recruited through a variety of methods. It uses matched sampling to survey respondents: The provider first draws a random sample from a large synthetic representative sampling frame, finds respondents that match the sampled individuals from its panel, and invites them to take a survey. For details and validation, see ?. For this particular sample, panelists also provided de-identified access to their web browsing activity via passive metering software installed voluntarily on their computers. The software, called Wakoopa, can be uninstalled at any point and captures visited web URLs independent of the type of browser or browser-specific privacy settings.[1] At the time this data was made available in March 2015, YouGov had recruited 1,392 individuals to the web tracking panel, which is currently marketed as YouGov Pulse. Our data comes from ?. See ?? for a comparison between demographics and population benchmarks on some key variables.

The passive metering component of this particular opt-in panel adds a layer of selectivity to the sampling process. Thus, to address concerns about the non-probability sampling procedure, we use raking weights to account for under- or overrepresentation along the following dimensions: age, gender, race, party identification, and region. We obtain population benchmark estimates from the U.S. Census (resident population, 2014) and proportions of party identifiers from Pew (2014). Raking takes into account the marginal, not the joint, distribution of the relevant demographic variables used to estimate the weights.

---

[1]Wakoopa does not save passwords or financial transactions, and personally identifying information is screened out by the survey provider.

Measuring Consumption of Adult Content

For both comScore and YouGov, we only observe data from a single machine per person. Our analyses should hold if people exhibit similar consumption patterns across devices. If that is too implausible an assumption, then we must decide on the direction of error and how it affects our analyses. We think it is likely that people would be less likely to search for pornography from machines on which they have installed these passive monitoring software (though the data are de-identified). If that is so, our estimates are a lower bound of net consumption of pornography per machine. As number of devices per person are increasing, all these numbers need to be adjusted. Next, is measurement error correlated with ideology? We have little reason to expect that, but we have no capacity to check if it is true. Thus, for current purposes, we assume that it is so.

We code pornographic content at the domain level. For both YouGov and comScore, we measure ideology using shallalist, DMOZ, and two machine learning classifiers. For YouGov data, we also present data based on classification from Wakoopa.

Our first classifier is based on just the domain name and domain suffix. In particular, we use a calibrated keyword classifier. The features of the model are whether any of the following keywords are present in the domain name:

cumshot, dildo, anal, adult, porn, mature, sex, xx, bbw, slut, whore, tits, titty, titties, pussy, sperm, gay, cheat, booty, ebony, asian, brazilian, fuck, cock, cunt, lesbian, shemale, boob, naughty, fatty, bitch, granny, jizz, faggot, horny, bukakke, bdsm, vagina, smut, x-rated, lusty, erotic, cunnilingus, blowjob, panty, hentai, latex, fetisch, fetish, erotik, bondage, naked, strip, teen, stocking, coitus, deprav, tube, perverse

For the 2004 comScore data, this gave about 140k domains that potentially carry pornographic content. We compare this list to the approximately 850k porn domains in the

shallalist. This leaves us with a list of 68k domains with uncertain status. Use one of the many URL classification APIs. Using Trusted Source API, we get about 20k porn, and 48k non-porn. This gives us the lower bound of adult domains.

To estimate the false positives, we take a large random sample of 10,000 unique domains and compare results from our classifier to results from a commercial API by McAfee. This serves are our estimate of the false positive rate. For comScore data, our false negative rate is 5%. False positive rate is close to 0. All of this ignores pornography available via more conventional channels. For instance, a large amount of pornography is consumed on sites like Tumblr.

0.1 Measuring Partisan Identification

With the YouGov survey, we measured ideology and partisanship using the conventional survey measures. In the comScore data, we do have data on partisan identification. We substitute with a validated model. We train our model on the YouGov data, predicting partisan identification using socio-demographics coded to match comScore (3-point party identification, gender, education, income, age, race, and ideology), and passively observed visits to site domains (excluding adult, rewards, and social networking sites). We aggregate the number of visits to each domain in the data by individual respondent, excluding domains with more than 100 visits total in the data. We estimate a cross-validated ridge (L2 penalized) regression with a total of 5,651 features. The model has a high degree of accuracy (92%).

Results

Figure 1: Consumption of Pornography by Ideology

4

Table 1: The effect of ideology on four separate dependent variables measuring pornography consumption.

	Number of visits	Pct. visits	Total time (seconds)	Pct. time
	(1)	(2)	(3)	(4)
Ideology: Conservative	−0.60*	0.001	−929.60	0.001
	(0.33)	(0.004)	(759.29)	(0.004)
Ideology: Don't know	−1.02**	−0.005	−862.56	0.001
	(0.44)	(0.01)	(1,068.65)	(0.01)
Ideology: Liberal	−0.64**	−0.004	−1,228.61	−0.003
	(0.30)	(0.004)	(771.12)	(0.004)
Ideology: Very conservative	−0.93**	−0.004	−1,265.14	−0.002
	(0.39)	(0.004)	(805.47)	(0.004)
Ideology: Very liberal	−0.04	0.01*	1,723.00**	0.01***
	(0.28)	(0.004)	(856.46)	(0.004)
Age: 25-44	−0.55**	−0.004	−2,024.46**	−0.01***
	(0.27)	(0.004)	(885.57)	(0.004)
Age: 45-64	−1.69***	−0.01*	−3,347.91***	−0.01***
	(0.35)	(0.004)	(907.86)	(0.004)
Age: 65+	−1.69***	−0.01**	−2,415.55**	−0.01***
	(0.41)	(0.005)	(1,035.71)	(0.005)
Gender: Male	2.39***	0.02***	3,448.81***	0.02***
	(0.31)	(0.002)	(515.79)	(0.002)
Married	−1.31***	−0.01***	−2,120.98***	−0.01***
	(0.28)	(0.003)	(546.76)	(0.003)
Constant	4.54***	0.01***	4,429.11***	0.02***
	(0.37)	(0.004)	(894.76)	(0.004)
N	1,367	1,367	1,367	1,367
Adjusted R^2		0.07	0.06	0.07

*p < .1; **p < .05; ***p < .01
Model 1: Quasi-Poisson; Models 2-4: OLS. Standard errors in parentheses.
The base category for ideological self-placement is "independent."
All models use weights raked to population by age, gender, race, party ID, and region.

Discussion

Supporting Information

SI 1 Comparing YouGov Sample to Population Benchmarks

	Unweighted	Weighted	Population
18-24	0.03	0.13	0.13
25-44	0.30	0.34	0.34
45-64	0.59	0.34	0.34
65+	0.07	0.19	0.19
female	0.56	0.51	0.51
male	0.44	0.49	0.49
asian	0.04	0.05	0.05
black	0.09	0.12	0.12
hispanic	0.05	0.14	0.14
other	0.07	0.05	0.05
white	0.76	0.64	0.64
ind	0.28	0.39	0.39
dem	0.37	0.32	0.32
oth/dk	0.12	0.06	0.06
rep	0.23	0.23	0.23
midwest	0.24	0.22	0.22
northeast	0.18	0.18	0.18
south	0.34	0.37	0.37
west	0.24	0.23	0.23
< $20,000	0.16	0.14	0.13
>= $120,000	0.09	0.10	0.11
$100,000-119,999	0.06	0.06	0.14
$20,000-39,999	0.25	0.25	0.17
$40,000-59,999	0.22	0.23	0.08
$60,000-79,999	0.14	0.15	0.20
$80,000-99,999	0.08	0.06	0.17