Deep Tags: Toward a Quantitative Analysis of Online Pornography

Antoine Mazieres, Mathieu Trachman, Jean-Philippe Cointet, Baptiste Coulmont, Christophe Prieur

▶ **To cite this version:**

Antoine Mazieres, Mathieu Trachman, Jean-Philippe Cointet, Baptiste Coulmont, Christophe Prieur. Deep Tags: Toward a Quantitative Analysis of Online Pornography. 2014. hal-00937745v1

HAL Id: hal-00937745

https://hal.science/hal-00937745v1

Preprint submitted on 28 Jan 2014 (v1), last revised 11 Aug 2014 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Deep Tags

Toward a Quantitative Analysis of Online Pornography

Antoine Mazières (SenS-INRA)

Mathieu Trachman (INED)

Jean-Philippe Cointet (SenS-INRA)

Baptiste Coulmont (Université Paris 8)

Christophe Prieur (LIAFA, Université Paris Diderot)

Abstract

The development of the web has increased the diversity of pornographic content, and at the same time the rise of online platforms has initiated a new trend of quantitative research that makes possible the analysis of data on an unprecedented scale. This paper explores the application of a quantitative approach to publicly available data collected from pornographic websites. Several analyses are applied to these digital traces with a focus on keywords describing videos and their underlying categorization systems. The analysis of a large network of tags shows that the accumulation of categories does not separate scripts from each other, but instead draws a multitude of significant paths between fuzzy categories. The datasets and tools we describe have been made publicly available for further study.

Keywords: online pornography; computational social sciences; sexual categories; network analysis

Introduction

> ‘The purpose of these keywords rests not upon their descriptive powers, but in the potential of naming. Naming creates both the symbology and the actuality of the world.’
>
> Lisa Z. Sigel 2000

When Linda Williams compared different kinds of pornography, revealing a proliferation of ‘diff’rent strokes for diff’rent folks’ (Williams 1992), she shed light on both historical and political phenomena. Indeed, during the 1970s there was a shift from a dominant male audience for pornography (Kendrick 1987) to diversified publics, along with the appropriation and staging of new desires. This ongoing diversification has been a central aspect of contemporary pornography, though it has been relatively unexplored.

Recently, this trend has been further amplified in line with a more general diversification of information sources and content, fostered largely by the development and democratization of the web and of media editing tools (Shirky 2008; Weinberger 2007). These have opened up niches for producers and broadcasters targeting a wide range of specific sexual desires (Williams 2004). The development of user-generated content has also contributed to the blurring of boundaries between amateur and professional, mainstream and alternative, and has permitted a variety of fantasies to be showcased (Paasonen 2010).

However, this proliferation has not been accompanied by a study of its dynamics. In Williams’ early article, sadomasochistic, homosexual and bisexual pornographies are taken to illustrate the gap between the norm and ‘perversity’, without taking into account the new interactions between categories that stem from their coexistence. It is the specificity of niches rather than the relations between them that is explored; for example, the appearance of new fantasies and their social background (Williams 2004), or the development of alternative pornographies (Jacobs et al. 2007; Taormino et al. 2013). But online pornography triggers new questions and internet activity provides logs of users’ activity, allowing quantitative analysis on an

unprecedented scale. Traces left by billions of users give us cultural snapshots of tastes and, more importantly, they enable researchers to look for structures and patterns in the evolutionary dynamics of practices adopted by a significant and growing proportion of the human population. As Hendler et al. note (2008) 'A large-scale system may have emergent properties not predictable by analyzing micro technical and/or social effects'. This opens the way for a 'computational social science' (Lazer et al. 2009), drawing on skills from various disciplines for processing computations on huge corpuses and interpreting their results with accuracy. This approach has been applied to many fields of inquiry, such as language dynamics (Lieberman 2007), evolution of science (Chavalarias and Cointet 2013), culture (Michel 2011), social networks (Easley and Kleinberg 2010), and epidemic forecasting (Ginsberg et al. 2008).

The availability of data from online platforms makes pornography a good candidate for such an approach. By collecting data on thousands of videos from two main pornographic platforms, we collected a large dataset of pornographic keywords and the relationships between them (where links exist between keywords that have been applied to the same videos). Our study focuses on categorization rather than consumption practices (Attwood 2005; Bozon 2012; Wright 2013), porn production (Edelman 2009) or the images themselves. The fact that the keywords are not randomly distributed means that they represent elementary atoms of information. If we were to postulate that *words inform sexuality* (Sigel 2000), our research explores the possibility that *porn tags inform pornography*.

Our hypothesis is that classification is not an organization of separated and hierarchical categories, as a Durkheimian perspective would suggest (Durkheim and Mauss 1901). It is not reducible to a virtuous circle, with practices and categories reinforcing each other and certifying the 'good' sexuality of those who are only heterosexual, monogamous, vanilla, and so on, as described by Rubin (2011). However, it does not follow that classification is anomic. In our datasets, discrete categories are related to each other and the whole system of relations exhibits a 'fuzzy logic'. The accumulation of categories does not separate fantasies from each other, but permits flow from one fantasy to another and draws thousands of paths corresponding to more and more precise desires. The proliferation of pornographic

categories does not only add minor fantasies to major fantasies; it shows how hegemonic desires provide a path to other desires, and how these other desires can be subsumed in hegemonic ones.

Several studies have applied quantitative schemes to traces from online pornography. Amanda Spink et al. (2004) analyzed the logs of two former web search engines for the year 2001 and identified the frequency of sexual queries within the whole corpus of web-search, along with the most frequent terms associated with them. The proportion of specific queries for illegal pornography such as child pornography in peer-to-peer networks, has also been studied (Latapy et al. 2013). In addition, general case studies with weblogs from several networks have been presented with collateral analysis of porn use. For instance, Berker (2002) analyzes a German university network and makes some observations about the volume and characteristics of porn-related traffic with respect to the network as a whole. A similar, more extended application of this approach can be found in the work of Ogas and Gaddam (2012) who analyzed 400 million search-engine queries in order to unveil the ‘billion wicked thoughts’ of its users.

In this article we present the methods used to acquire our datasets and their main characteristics, and go on to focus on the underlying classification systems and the structural differences they imply. Online content categorization has been the focus of many studies of online interaction and collaboration (Guy and Tonkin 2006; Cattuto et al. 2009). We recall one of their major structural elements, namely the highly skewed distribution of the categories: a large proportion of items are covered by a very small number of almost universal categories, while a long tail of more specific categories still gather a large variety of content (Anderson 2006). This phenomenon encourages great diversity in content and induces the development of niches (Brynjolfsson 2006). We explore various methods for analyzing categories, ranging from frequency measurement to network analysis, in order to reveal the diversity behind hegemonic categories, and the means by which the interactions within them are assembled into niches.

1. Classifying One's Desire: Dataset Acquisition And Description

Online porn is available in numerous forms. Because of their small size, plain text stories, picture galleries and comics were probably the first types of porn content to be widely diffused on the web. Audio and video files came later, with video the main medium during the 2000s, largely due to the wider availability of broadband internet connections and better streaming technologies which have enabled us to view, upload and host videos easily. However, video-hosting platforms are in competition with other kinds of services (Ogas and Gaddam 2012) which enable direct interaction between pornographic actors and viewers. For example, LiveJasmin.com, a webcam-based interaction platform, is ranked as the 3rd most visited website in the adult category[i]. Webcam communities broadcast unstructured content - often streamed video and chat - which is unarchived and has little metadata. Despite the importance of this growing medium of online pornography, the lack of structure in the data means it is outside the scope of our study. Video-hosting platforms, on the other hand, present well-structured data. Every video belongs to a page, with a specific URL, a list of associated keywords and various other metadata such as the number of views, upload date, comments, votes, descriptions, and so on. This information is publicly available to any user and the method we used to collect our data differs from that used by a regular user only in its systematic approach.

1.1 Datasets acquisition and description

According to several website popularity rankings[ii], we identified the two most popular pornographic video hosting platforms - XNXX and XHamster. We created a dedicated computer program to carry out the navigation and data collection tasks required to gather the metadata for all available videos on both websites without downloading any videos. The datasets are available online[iii] and released under a Creative Commons License[iv]. As shown in Tables 1 and 2, a variety of data is attached to each entry. The last column indicates how much of the dataset's entries are provided with the data described in each row.

XNXX and Xvideos[v] domains are the oldest among the most popular porn platforms, dating from 1997. In July 2013, the websites claim to host more than 3.5 million

videos. We gathered information for 1,166,278 videos that were uploaded before March 2013. XNXX releases very little data about the videos it hosts. As shown in Table 1, only the title, keywords and comments are available to the public. Information about uploaders and the number of views is hidden or not logged by the platform maintainers.

Data ID	***Description***	***% of the dataset***
title	Title of the video	100 %
nb_comments	Number of comments posted on this video	99 %
tags	List of the keywords associated with this video	93 %

Table 1: Description of XNXX dataset

Our interest in this dataset lies primarily in its tags. When someone uploads a video, they can attach any number of keywords to their file. These keywords are meant to describe the video and highlight its specificities in order to help the user find it more easily, by anticipating the words used in a search query targeting this content. By allowing uploaders to index their videos with numerous keywords, XNXX possesses a corpus of over 70,000 tags. Among the most common pornographic platforms, XNXX is the only one to have such a corpus of descriptive keywords.

XHamster is a recent platform dating from 2007 and probably for this reason hosts fewer videos. All of the videos can be accessed, and our dataset includes all the videos hosted by the platform since its creation and still available when we collected the data in February 2013. This represents 786,121 entries in the format described in Table 2.

Data ID	*Description*	*% of the dataset*
title	Title of this video	100 %
upload_date	Day when the video was uploaded	99 %
channels	List of the keywords associated to this video	99 %
Nb_views	Number of times this video has been displayed	99 %
Nb_votes	Number of users who voted on this video	99 %
runtime	Length of the video in seconds	99 %
uploader	Anonymized identifier of the uploader's username	95 %
nb_comments	Number of comments posted on this video	92 %
description	Description attached to this video	48 %

Table 2: Description of XHamster dataset

The presence of a timestamp on 99% of the videos permits analyses of changes through time[vi]. To avoid taking incomplete years into account while considering metadata evolution, years 2007 and 2013 are omitted. An anonymized identifier links the uploader to their video-clips. This permits us to track the repetition of videos among uploaders and the relations between uploaders with specific content categories or videos characteristics (e.g. runtime, comments, views)[vii].

As two of the most important pornographic platforms, XNXX and XHamster offer a representative sample for studying online pornography. Moreover, the structure of their data is significantly different which makes them amenable to a comparative approach.

1.2 Categorization Systems

Tags, categories and keywords are similar words to speak about *semantic descriptors*. They are fundamental elements of the contemporary web: they sort

content into menus and lists. They are the basis of the algorithms which allow content to be indexed in such a way as to improve the searching and browsing experiences of users. On pornographic platforms, keywords may describe practices (‘BDSM’, ‘blowjob’), ethnic or cultural characteristics of actors (nationalities, geographical region, skin colour, religion), places (bus, bedroom, public places), devices (bed, dildo), filming techniques (‘point of view’, ‘hidden’, ‘hd’) and so on (Tan Hoang 2004; Attwood 2010). The keywords define the degree of semantic diversity available to uploaders in their content descriptions, and to viewers in their search queries.

On both XNXX and XHamster, videos are categorized by their uploaders. However, the platforms have different categorization systems. XHamster has a traditional top-down system which limits uploaders to pre-determined categories for characterizing their content, and viewers correspondingly only have these categories available for identifying content. This is the most common approach to categorization in pornographic platforms, most of them providing a similar list of ‘classic’ categories. XNXX has a bottom-up approach, letting uploaders choose their own words to index their videos, resulting in a list of more than 70,000 so-called ‘tags’. This system offers greater semantic variety to the viewers, facilitating the emergence of keywords and their combinations.

The difference between top-down categories and bottom-up tags is characteristic of changes in classification strategies and practices in the digital era (Bowker and Leigh Star 1999; Weinberger 2007). The latter - known as ‘folksonomy’ - is a key feature in the development of content diversity and, in our case, in the tracking of contemporary porn diversification (Attwood 2007). The substantial difference in the range of semantic possibilities for uploaders and viewers impacts the number of dimensions indexed by the platforms and is therefore observable in our study.

However, despite the two platforms having different categorization systems, there are some strong similarities between the datasets, which suggests a possible generalization to other pornographic platforms. One structural similarity is that whatever the number of categories available, a very small number of tags allows one to access most of the content. For instance, on XNXX, the top 5% of the most

popular tags covers more than 90% of the videos. On both XHamster and XNXX, the most frequent categories, respectively amateur and blowjobs, targets 30% of all entries. To further explore the datasets beyond the identification of the few dominant widespread categories, we designed several other methodological tools.

2. From Frequency To Network

Behind this structure lies a 'long tail' of less common sexual scripts and descriptors, calling for finer-grained approaches. We first rank tag frequencies by their occurrence in titles, or using alternative methods. Then, taking into account the highly skewed distribution of tags, we shift our focus to the relationships between them. Network analyses of these relationships allow us to monitor the dominance of certain tags, revealing porn semantics diversity and the niches within its network.

2.1 The Hegemony Of High Frequencies

2.1.a Word Frequency in titles

All the videos possess one title describing their content. Some recurring archetypes (such as 'boss', 'secretary', 'maid', 'brother's best friend', and so on) can be identified in the datasets. The words 'mom' or 'mother' are present in 37 of the 100 most seen videos in XHamster. Therefore, while our study focuses on more structured aspects such as categories, we have released a tool[viii] for plotting and comparing word frequencies over time in video titles from the XHamster dataset (Fig. 1).

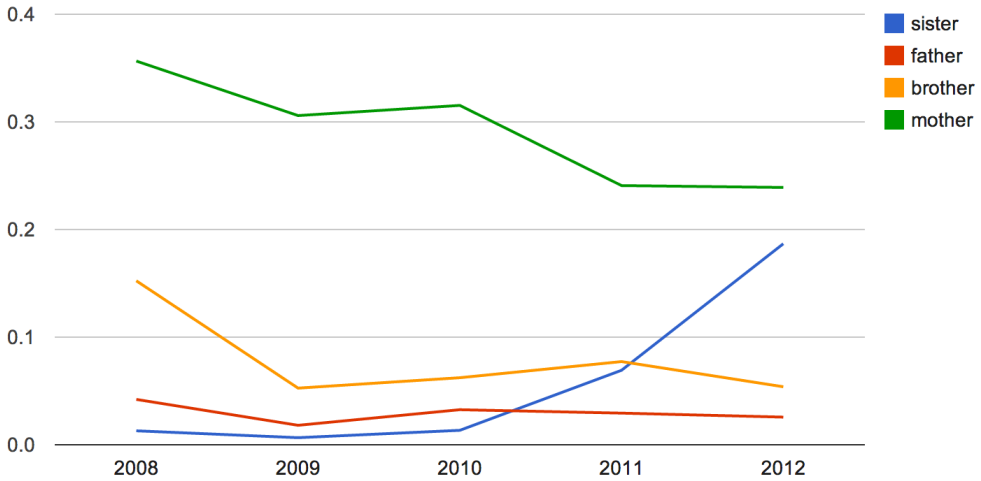

Figure 1: Example of query

The fact that titles are unstructured sequences of characters poses challenges for conducting a systematic analysis. Spelling and typing errors, abbreviations, uses of plural and conjugated forms can all result in significant biases. For word frequencies in XHamster's titles, our algorithm strips out dashes and catches any occurrence of the query in the title, for example, 'blow' catches 'blowing', 'blowjobs', and so on while leaving biases from typing errors ('blwjob') and abbreviations ('bj') unhandled. In this example, adding typing errors and abbreviations increases the number of blowjob videos by 16%.

2.1.b Category Frequencies

For tag frequencies in XNXX, our algorithm only catches the specific instance of the query, which means 'blowjob' will only catch the tag 'blowjob' (case-insensitive). By considering [blowjob(s), blowing, bj, blow(s), blow-job(s), blowwjob, blwjob] as variants of 'blowjob', we increase the number of videos considered in XNXX by 5%. The bias induced by typing errors and abbreviations is thus significantly lower than for word frequencies in titles, even though our algorithm catches no variants. This phenomenon is induced by *folksonomies* (Halpin et al 2009; Cattuto et al 2009) where uploaders tagging their videos make a greater effort to use the most common descriptor than when they are writing titles.

We can rank categories by their frequency of occurrence, that is, for each tag, the number of videos having that tag (most videos have several tags). The top keywords represent the descriptors from which most of the videos can be accessed. If they illustrate strong practices or cultural trends, they may also overlap with other categories and get their dominant position from the transversality or generality of the concept they refer to. For example, 'amateur' and 'blowjob' do not exclude many other categories, such as those derived from sexual practices, nationalities, ethnic groups, scenarios, and so on. Adding other dimensions to the ranking by occurrences allows us to highlight interesting properties of pornographic content descriptors.

Occurrences		**Popularity**	**User reaction**	
XHamster	*xnxx*	*XHamster*	*XHamster*	*xnxx*
Amateur	blowjob	Grannies	Cuckold	muslim
Men	hardcore	Old+Young	Midgets	hijab
Teens	amateur	Korean	Grannies	arabic
Hardcore	teen	Matures	Bisexuals	step
Blowjobs	cumshot	Arab	Strapon	tribadism
Anal	anal	Midgets	Cream Pie	girlontop
Big Boobs	brunette	Massage	Shemales	arabe
Masturbation	blonde	Swingers	Matures	cody
Matures	pussy	Italian	Old+Young	cumglass
Cumshots	sex	Turkish	German	sister

Table 3: Various ranking methods over tags, top 10.

Popularity ranking is only available for XHamster and reveals categories by the number of views generated by all videos in a given category, 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 categories may point to content for which demand surpasses what is offered by uploaders.

User reaction ranking tends to increase the average number of comments per video of the given category. This uncovers viewers' reactions and interactions around the video's content. Without reading the actual comments, it is difficult to determine whether, for example, the reactions are simply descriptive or not. However, some videos may trigger comments and discussion.

Table 3 only provides the top ten tags for each of the rankings, but we have released the dataset for all tags to permit further studies to be carried out[ix]. Ranking tags allows us to isolate the various properties of specific porn content descriptors compared to the others. However this focus tends to mute the high number of tags that, while not among the most frequent, still have significant levels of popularity in

terms of number of videos. Taking into account tags co-occurrences provides a far finer-grained tool for analysis, which we detail below.

2.2 Porn Semantics As A Network

2.2.a Link over-representations: 'blowjob' doesn't make it 'funny'

The majority of videos in our dataset are attached to more than one category. If we consider the presence of several categories for the same videos as a link between each of these keywords, then we can build a global 'semantic' network. Categories are nodes connected through an edge (link) when two categories are significantly 'close' to one another. Such an analytical framework, known as network analysis and coming from the study of social relationships (Scott & Carrington 2011), has become very popular in many fields (Easley & Kleinberg 2010, Newman, 2010).

As we have observed, tag frequencies are highly heterogeneous. This is the reason why we cannot simply rely on a raw count of co-occurrences to assess the relation strength between two tags. While we are aiming at describing only preferential relationships, very frequent tags such as 'amateur' or 'blowjobs' would obviously co-occur with any other tag. A measure of proximity must be defined for capturing how much the actual number of co-occurrences deviates from the theoretical value one would expect with no correlation between tags[x]. By doing so, we focus on edges between strongly connected tags.

As an illustration, 'midgets' - a low frequency category in XHamster - is present ten times more than expected in all videos having the tag 'funny'. This indicates a strong relation between these two categories and tells us that it is highly likely that midgets appear mainly to fulfil a 'funny' aspect of the scene. The fact that 'midgets' appears more with 'blowjobs' than with 'funny' is *statistically expected* and therefore ignored, while the relation between 'midgets' and funny' is *unexpected*, and consequently highlighted in the network.

Given this methodology, we can look at link over-representation for each category without dominant categories swamping awareness of the strong and meaningful

symbolic associations between less frequent categories[xi]. Taking into account link over-representation reveals widely adopted symbolic associations between categories of the considered pornographic content.

These strong relations might illustrate obvious associations, such as tools or practices for a given behaviour, geographical region or ethnicity for a nationality, and so on. They allow more surprising observations when types of categories are mixed, for instance a nationality with an object or a practice. To illustrate such associations, we took the administrative and political entity named by categories (which we considered to be the common chunk of cultural entities) and identified their privileged relations with other types of categories. Table 6 shows the three strongest links for all categories referring to a nationality.

Nationality categories	**3 most overrepresented associated categories**
Japanese	Asian, Massage, Bukkake
German	Vintage, Gothic, Grannies
French	Arab, Anal, Gangbang
British	Stockings, Bukkake, Celebrities
Russian	Babysitters, Old+Young, Teens
Indian	Arab, Asian, Emo
Brazilian	Latin, Anal, Black and Ebony
Italian	Celebrities, Vintage, Old+Young
Turkish	Arab, Funny, Celebrities
Czech	Spanking, POV, Old+Young
Thai	Asian, Massage, Squirting
Korean	Asian, Chinese, Hidden Cams
Chinese	Asian, Korean, Japanese
Swedish	Danish, Vintage, Gothic

Table 4: Example of link overrepresentation between categories (XHamster)

A video uploaded with a nationality category does not necessarily take place in the related country or show actors coming from it. It does not accurately inform us of a country's sexual practices, but rather serves as an indicator of how this nationality is staged in a pornographic context. These examples may be applied to the whole set of relationships between the categories to obtain more generalized, global conclusions.

2.2.b Porn semantic network

Figure 1 helps visualize the whole network obtained from the XHamster dataset. Only edges whose strengths are above a given threshold have been represented. An algorithm has automatically determined this threshold such that the final network is as sparse as possible but still composed of one unique connected component. We applied a community detection method, often referred to as Louvain algorithm (Blondel et al. 2008), to identify cohesive subsets of tags in the corpus. These 'clusters' gather densely connected tags which are relatively disconnected from the rest of the network and may form semantically coherent units. In Figure 1 each node is coloured according to the clusters to which it belongs.

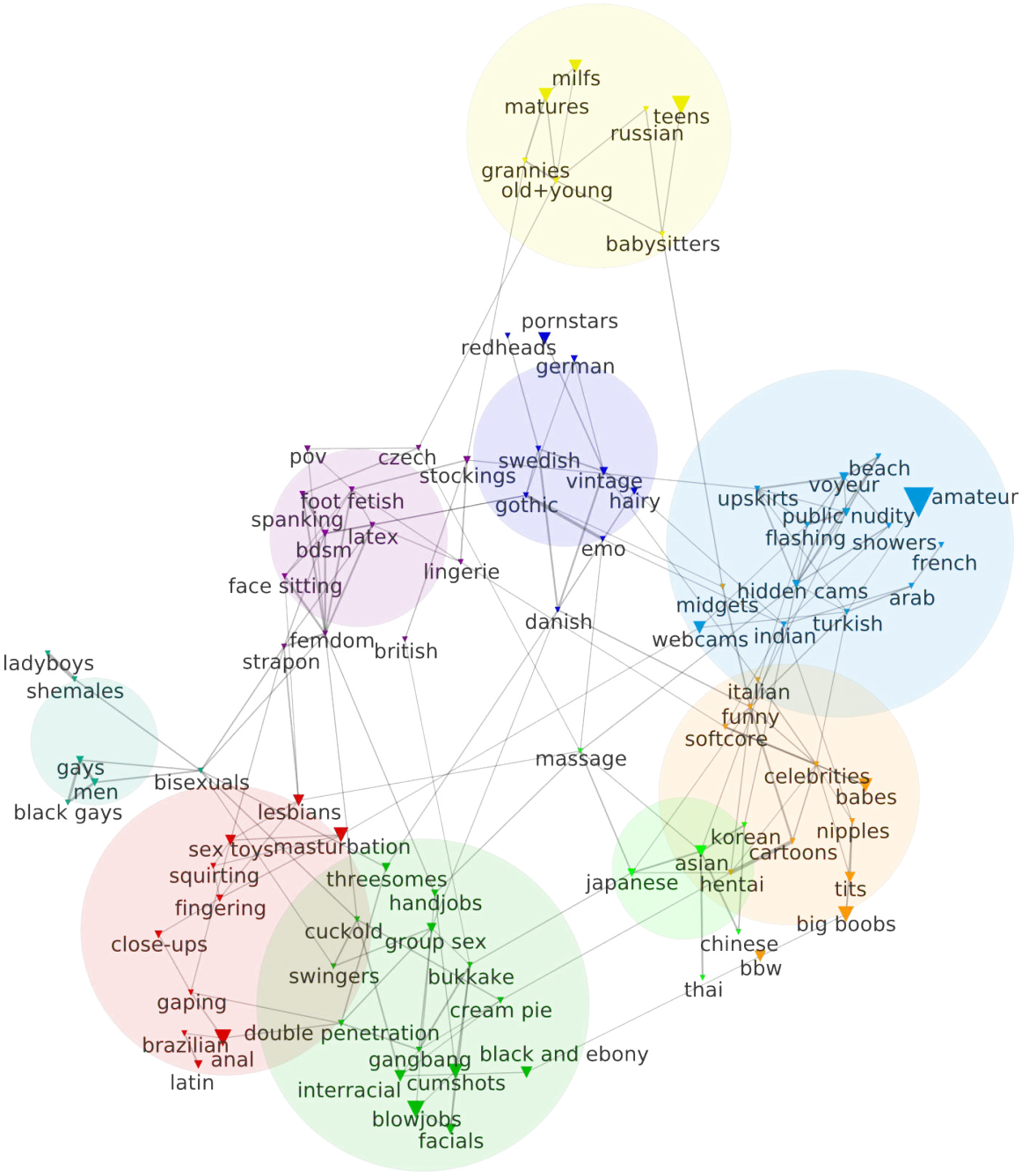

Figure 2: XHamster tag network

As well as the statistical measures available for network analysis, one can also sketch qualitative observations from visualization to characterize the network's structure and the relations between the nodes. Some clusters are highly thematic, referring to age (milfs, teens, matures, grannies), practices such as BDSM (latex, spanking, facesitting), context (beach, voyeur, flashing, public) or nationalities (thai, chinese, korean, asian). Other clusters are more heterogeneous and mix different types of keywords, such as 'blowjobs', 'black', 'ebony' and 'threesome'. The

presence of hubs between several clusters is another remarkable property, such as ‘massage’ or ‘Danish’ having links with many others clusters, strong enough to appear in this visualization.

Among many other possible assertions, it is worth noting the strong separation of the cluster containing the tags ‘gay’ and ‘transsexual’ from all other parts of the network. Indeed, it is connected to the rest of the network only through the tag ‘bisexual’ which constitutes a privileged bridge for any other co-occurrence. The position of the gay cluster strongly reinforces a division between heterosexuality and homosexuality by isolating the latter (Sedgwick 1990). As Halperin states (1995: 44), ‘Heterosexuality defines itself without problematizing itself, it elevates itself as a privileged and unmarked term’. This privilege makes heterosexuality subsume most sexual categories, e.g. what is not “heterosexual” must be defined. It therefore acquires more semantic influence upon the repertoire of desires and fantasies available on pornographic platforms. This isolation of ‘gays’ calls for a more general analysis of cases where some categories or groups of categories become to some degree peripheral to the network and constitute niches.

2.3 On Categories Nicheness And Datasets Limits

We observed on the previous network that some nodes have high degrees (that is, many links) and occupy relatively central positions in the network, while others are only connected to a few other tags and seem more peripheral in the general picture. To measure more rigorously such a property we designed a so-called *nicheness coefficient*. The nicheness coefficient is built upon the global matrix of mutual information between pairs of tags. We simply define the nicheness score of a tag as the sum of the preferential links connecting this tag to its relevant neighbours. The rationale behind such a measure is that tags with a ‘niche’ behaviour, that is tags compatible with only few other tags, will be connected by very strong edges. Conversely, tags that may be used in conjunction with any other tags are likely to have many weakly connected neighbours and a degree of distribution that is close to random, thus resulting in a very low nicheness score. Put differently, nicheness score also measures how much the probability of using a tag is dependent or not on the presence of other tags. If this probability remains largely unchanged with

different tag pairings, the tag nicheness score is low. If the presence/absence of another tag strongly increases/decreases (and vice-versa) the probability to observe a tag, then the tag has a higher nicheness score.

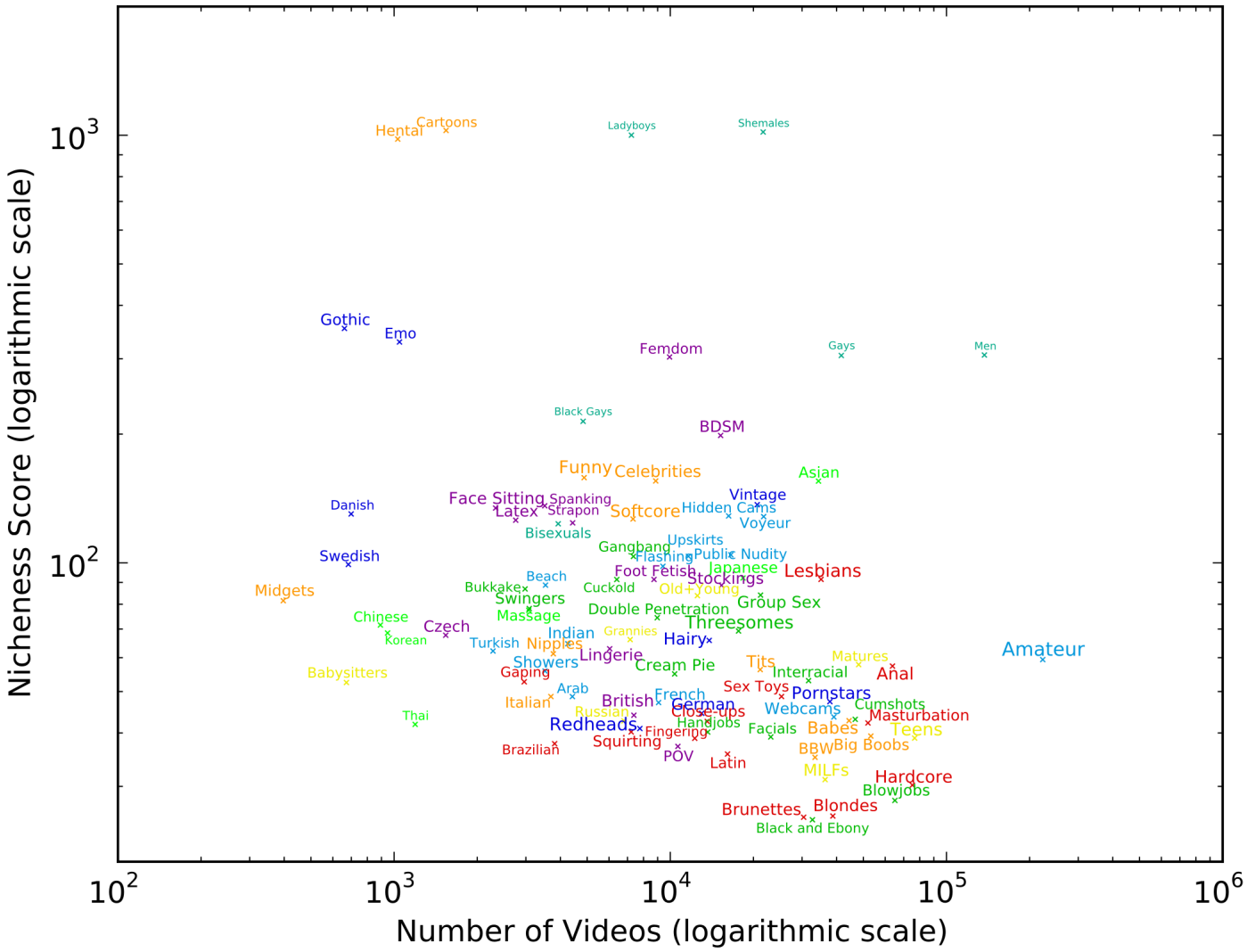

Figure 3: Nicheness of XHamster categories.

Figure 3 shows a scatter plot of the 92 XHamster channels according to frequency and nicheness. Label size scales with tag degree and node colours are consistent with Figure 1. We observe that ‘hentai’ and ‘cartoons’, although compatible with a respectable number of tags still have a very ‘biased’ distribution of co-occurrences, leading to one of the highest nicheness score. Similarly, ‘ladyboys’ and ‘shemales’ feature high nicheness score but have very low degrees (namely 1 and 3). It is

interesting to note that niche tags are not necessarily rare. ‘Men’ is among the 10 tags with the highest nicheness score and it is the second most frequent channel. Higher nicheness score corresponds to tags that target more specialized resources. In contrast, low nicheness score tags are compatible with many other tags, and therefore provide less certain and/or less fine-grained descriptions of the content.

This empirical measure of nicheness improves upon the usual descriptions of porn niches. The niches described in Williams (1992) are practices such as BDSM that are situated outside of Rubin’s virtuous circle and practices akin to perversions of vanilla sex, whereas the many niches of online porn are in a state of flux and stem from the mobilization of specialized resources. It is not shifts in which perversions are put on scene that form the basis of this specialization of niches, but rather specialization within major and minor sexual practices and identities (Penley 2004).

Online pornography consumers are unlikely to be immobile in the landscape of niches described by Figure 3. Some niches bring users to other niches; some of them might even attract newcomers, while others might repel viewers from porn. The paths of users within the search space should exhibit patterns relevant to understanding their ‘careers’ as porn consumers. Structured computer traces and other data from hundreds of millions of consumers would provide material to study pornography on an unprecedented scale. However, due to the fact that the traces left by users (mainly identification and geolocalization) on the platforms’ servers are possessed by the owners of the hosting sites and not publicly available, our dataset does not include data directly linked to users’ behaviours. Access to such data would extend our approach and shed light on the symbols linking niches through first-hand observation of users’ careers within this content.

Furthermore, tags can have different meanings in different contexts. Uses of porn categories greatly depend on national and geographical context. For example, the ‘Beurette’ (Arab girl in French) category is not understandable in isolation from an understanding of the French colonial past and postcolonial contemporary relationships, which produce young Arab girls as objects of desire for a white male gaze (Fassin and Trachman 2013). The potential nicheness of ‘Beurette’ in France could be compared to the mainstreamness of ‘Arab’ in North Africa or Middle East

regions. We could say the same thing for the apparently most transparent ‘gay’, whose application varies with the different meanings of heterosexual/homosexual binarism and with cultural contexts of moral, law and sexuality. Accessing geolocalized information would therefore help to contextualize different semantic elements within their cultural surroundings.

Conclusion

By focusing on publicly available data, this study sought to determine whether porn tags provide a way of informing research on pornography. It appears that such an approach does help us to shed light on the structural properties of porn tags so as to identify the widespread presence of dominant categories and to reveal diversity in the ‘long tail’ of less common sexual scripts. Beyond this general view of porn semantics, we analyzed its more discrete descriptors, involving specific users, and their privileged interactions with other words. These words and their specific layouts yield heterogeneous communities of practices, objects, actors and places which inform pornography.

Our goal, using a massively quantitative approach to these phenomena, was not only to measure dominant vs. underrepresented categories, but to look at categorization practices in pornography. By modeling and visualizing this data, we enabled qualitative assessments to be made of tags’ positions in networks and the links between categories, and therefore of how practices, nationalities, places and techniques are staged in the pornographic landscape. Large datasets and tools permit more statistical explorations and validation, but also allow a qualitative approach to be taken with respect to their numerical and visual outputs. A small-scale approach to large-scale results is likely to provide richer and more detailed information on specific communities and users.

Our study reverse-engineers users’ ‘tastes and colours’ through the analysis of platforms structure and uploaders behaviors. While highly relevant for both website maintainers and content diffusers in devising strategies to target users, users’ practices are not well understood because their traces are owned and kept by the websites. However, platform maintainers have carried out several initiatives[xii].

Beyond the obvious ‘buzz’ and ‘safe for work’ marketing strategies whose purpose is to encourage people to discover and discuss the existence of such and such platform, the data and related analyses are not verifiable. But these leaked user traces serve as evidence confirming the existence of this data in the hands of platform maintainers and their unexplored scientific potential. Allowing researchers to access this data would allow a wide range of possibilities for understanding how pornography is used and the aspects of human sexuality it represents.

Our interdisciplinary study presents the initial results of more long-term research that aims to articulate the possible contribution of large-scale quantitative methods to the theoretical and analytical frameworks provided by porn studies to understand pornographic contexts and actors. By making our datasets, analysis and tools publicly available, we hope to make this approach more accessible to those wishing to extend this approach and/or to focus more specifically on particular communities and practices, or on other aspects of porn.

References

Anderson C. 2006. *The Long Tail: Why The Future Of Business Is Selling Less Of More*. New York: Hyperion.

Attwood, F. 2005. ‘What Do People With Porn? Qualitative Research Into The Consumption, Use And Experience Of Pornography And Other Sexually Explicit Media’. *Sexuality & Culture* 9(2): 65-86.

Attwood, F. 2007. ‘No Money Shot? Commerce, Pornography And New Sex Taste Cultures’. *Sexualities* 10(4): 441-456.

Attwood, F. ed., 2010. *porn.com. Making Sense Of Online Pornography*. New York: Peter Lang.

Berker, T. 2002. ‘World Wide Web Use At A German University - Computers, Sex, And Imported Names: Results Of A Log File Analysis’. In: *Online Social Sciences*, edited by B. Batinic, U.D. Reips and M. Bosnjak, 365-382. Göttingen: Hogrefe.

Blondel, V.D., Guillaume, J.L., Lambiotte, R. and Lefebvre, E. 2008. ‘Fast Unfolding Of Communities In Large Networks’. *Journal Of Statistical Mechanics: Theory And Experim*ent 10: 10008.

Bowker, G. and Leigh Star, S. 1999. *Sorting Things Out: Classification And Its Consequences*. Boston, MA: MIT Press.

Bozon, M. 2012. 'Sexual Encounters And Sexual Practices: A Widening Repertoire'. In: *Sexuality In France. Practices, Gender & Health*, edited by N. Bajos & M. Bozon, 243-264. Oxford: The Bardwell Press.

Bronstein, C. 2011. *Battling Pornography: The American Feminist Anti-Pornography Movement, 1976-1986*. Cambridge: Cambridge University Press.

Brown, J. D. & L'engle, K. 2009. 'X-Rated: Sexual Attitudes And Behaviors Associated With U.S. Early Adolescents' Exposure To Sexually Explicit Media'. *Communication Research* 36(1): 129-151.

Brynjolfsson, E., Hu, Y.J. and Smith, M.D. 2006. 'From Niches To Riches: The Anatomy Of The Long Tail'. *Sloan Management Review* 47(4): 67-71.

Cattuto, C., Barrat, A., Baldassarri, A., Schehr, G. and Loreto, V. 2009. 'Collective Dynamics Of Social Annotation'. Proceedings Of The National Academy Of Sciences Of The United States Of America. 106(26): 10511-10515.

Chavalarias, D. and Cointet, J.P. 2013. Phylomemetic Patterns In Science Evolution - The Rise And Fall Of Scientific Fields. *Plos One*, 8(2): e54847.

Desrosieres, A. 1998. *The Politics Of Large Numbers. A History Of Statistical Reasoning*. Cambridge, MA: Harvard University Press.

Durkheim, E. and Mauss, M. 1901. *De Quelques Formes Primitives De Classification: Contribution À L'étude Des Représentations Collectives*. L'année Sociologique (1896/1897-1924/1925) 6: 1-72.

Easley, D. and Kleinberg, J. 2010. *Networks, Crowds, And Markets: Reasoning About A Highly Connected World*. Cambridge: Cambridge University Press.

Edelman, B. 2009. 'Red Light States: Who Buys Online Adult Entertainment?' *Journal Of Economic Perspectives* 23(1): 209-220.

Fassin E. and Trachman M. 2013. 'Voiler Les Beurettes Pour Les Dévoiler: Les Doubles Jeux D'un Fantasme Pornographique Blanc'. *Modern & Contemporary France* 21(2): 199-217.

Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S. and Brilliant, L. 2008. 'Detecting Influenza Epidemics Using Search Engine Query Data'. *Nature* 457(7232): 1012-1014.

Granovetter, M.S. 1973. The Strength Of Weak Ties. *American Journal Of Sociology* 78(6): 1360-1380.

Guy, M. and Tonkin, E. 2006. Tidying up tags. *D-lib Magazine 12*(1): 1082-9873.
Halperin, D.M. 1995. *Saint Foucault. Toward A Gay Hagiography*, New York: Oxford University Press.
Halpin H., Robu V. and Shepherd, H. 2009. 'Emergence Of Consensus And Shared Vocabularies In Collaborative Tagging Systems', *ACM Transactions On The Web* 3(4): 1-34.
Hendler, J., Shadbolt, N., Hall, W., Berners-Lee, T. and Weitzner, D. 2008. 'Web Science: An Interdisciplinary Approach To Understanding The Web'. *Communications Of The ACM* 51(7): 60-69.
Jacobs, K. 2007. *Netporn: DIY Web Culture And Sexual Politics*. New York: Rowman & Littlefield.
Kendrick, W. 1987. *The Secret Museum: Pornography In Modern Culture*. Berkeley: University Of California Press.
Jacobs K., Janssen M. and Pasquinelli, M. eds., 2007. *Click Me: A Netporn Studies Reader*, Amsterdam, Institute Of Network Cultures.
Latapy, M., Magnien, C. and Fournier, R. 2013. Quantifying Paedophile Activity In A Large P2p System. *Information Processing & Management* 49(1): 248-263.
Lazer, D. et al. 2009. *Life In The Network: The Coming Age Of Computational Social Science. Science* 323(5915): 721-723.
Lieberman, E., Michel, J.B., Jackson, J., Tang, T. and Nowak, M.A. 2007. Quantifying The Evolutionary Dynamics Of Language. *Nature* 449(7163): 713-716.
Mandelbrot B. 1957. 'Étude De La Loi D'estoup Et De Zipf: Fréquences Des Mots Dans Le Discours'. In: *Logique, Langage Et Théorie De L'information*, edited by L. Apostel, B. Mandelbrot and A. Morf, 22-53. Paris: Presses Universitaires de France.
McPherson, M., Smith-Lovin, L. and Cook, J.M. 2001. 'Birds Of A Feather: Homophily In Social Networks'. *Annual Review Of Sociology* 27: 415-444.
Michel, J.B., Shen, Y.K., Aiden, A.P., Veres, A., Gray, M.K., Pickett, J.P. and Aiden, E.L. 2011. 'Quantitative analysis of culture using millions of digitized books'. *Science* 331(6014): 176-182.
Newman, M. 2009. *Networks: An Introduction*. Oxford: Oxford University Press.
Ogas, O. and Gaddam, S. 2011. *A Billion Wicked Thoughts: What The Internet Tells Us About Sexual Relationships*. New York: Penguin.
Paasonen S. 2010. 'Labors Of Love: Netporn, Web 2.0 And The Meaning Of Amateurism'. *New Media & Society* 12(8): 1297-1312.

Penley, C. 2004. ‘Crackers And Whackers. The White Trashing Of Porn’. In: *Porn Studies* edited by L. Williams, 309-320. Durham: Duke University Press:

Rubin G. (2011). *Deviations. A Gayle Rubin Reader*. Durham: Duke University Press.

Scott J. and Carrington P.J. eds., 2011. *The Sage Handbook Of Social Network Analysis*. London: Sage.

Sedgwick, E.K. 1990. *Epistemology Of The Closet*. Berkeley: University Of California Press.

Segal L. and McIntosh, M. eds., 1992. *Sex Exposed. Sexuality And Pornography Debate*, London: Virago Press.

Shirky, C. 2008. *Here Comes Everybody: The Power Of Organizing Without Organizations*. New York: Penguin.

Sigel, L.Z. 2000. ‘Name Your Pleasure: The Transformation Of Sexual Language In Nineteenth-Century British Pornography’. *Journal Of The History Of Sexuality* 9(4): 395-419.

Spink, A., Koricich, A., Jansen, B.J., & Cole, C. 2004. ‘Sexual Information Seeking On Web Search Engines’. *Cyberpsychology & Behavior* 7(1): 65-72.

Tan Hoang, N. 2004. ‘The Resurrection Of Brandon Lee: The Making Of A Gay Asian American Porn Star’. In: *Porn Studies* edited by L. Williams, 223-270. Durham: Duke University Press.

Taormino, T., Parreñas Shimizu, C., Penley, C. and M. Miller-Young eds., 2013. *The Feminist Porn Book. The Politics Of Producing Pleasure*, New York: The Feminist Press.

Trachman M. 2013. *Le Travail Pornographique*: *Enquête Sur La Production De Fantasmes*. Paris: La Découverte.

Warner, M. 1999. *The Trouble With Normal: Sex, Politics And The Ethics Of Queer Life*. Cambridge, MA: Harvard University Press.

Weinberger D. 2007 *Everything Is Miscellaneous: The Power Of The New Digital Disorder*. New York: Henry Holt Company.

Williams L. 1992.‘Pornographies On/Scene, Or Diff’rent Strokes For Diff’rent Folks’. In: *Sex Exposed: Sexuality and the Pornography Debate* edited by L. Segal and M. McIntosh, 233-265. London: Virago.

Williams, L. 2004. ‘Porn Studies: Proliferating Pornographies On/Scene: An Introduction’. In: *Porn Studies* edited by L. Williams, 1-23. Durham: Duke University Press:

Wright, P.J. 2013. 'U.S. Males And Pornography, 1973-2010: Consumption, Predictors, Correlates', *Journal Of Sex Research* 50(1): 60-71.

[i] http://www.alexa.com/topsites/category/Top/Adult
[ii] Alexa and Netcraft rankings, accessed in August 2013.
[iii] Alexa and Netcraft rankings, accessed in August 2013.
[iv] http://pornstudies.sexualitics.org/#dataset is https://creativecommons.org/licenses/by/3.0/deed.en_US
[iv] https://creativecommons.org/licenses/by/3.0/deed.en_US
[v] XNXX and Xvideos are two interfaces to the same corpus of videos.
[vi] For instance, the average runtime has been multiplied by 7. Also, runtime varies a lot between categories (23 minutes for *double penetration* and 4 minutes for *men*).
[vii] Our dataset covers the contributions of 90,000 uploaders, half being one-time uploaders only, representing only 10% of the videos.
[viii] http://porngram.sexualitics.org/
[ix] http://pornstudies.sexualitics.org/#catrank
[x] More precisely, denoting n(i) the number of videos featuring tag i and n(j) the number of videos in which j is mentioned. The edge strength is defined as the ratio between observed and theoretical values of videos using both i and j which can be computed as s(i,j)=(n(i,j)N)/(n(i)n(j)) with N = total number of videos.
[xi] Full dataset available at http://pornstudies.sexualitics.org/#link
[xii] PornMD released an interface to explore the ten most queried tags by country: http://www.pornmd.com/sex-seach
Pornhub, since June 2013, regularly release data and exploration tools on their data: http://www.pornhub.com/insights/
TorrentFreak looked at porn queries coming from specific countries: http://torrentfreak.com/priests-watch-dvd-screeners-while-pirates-download-filth-in-the-vatican-130407/