Noname manuscript No.
(will be inserted by the editor)

Adult Content Consumption in Online Social Networks

Mauro Coletto · Luca Maria Aiello · Claudio Lucchese · Fabrizio Silvestri

Final version available at https://doi.org/10.1007/s13278-017-0449-y

Abstract Users in online social networks naturally organize themselves into overlapping and inter-linked communities that are formed around common identity or shared topical interests. Some communities gather people around specific *deviant behaviors*, conducts that are commonly considered inappropriate with respect to the society's norms or moral standards such as drug use, eating disorders, and pornographic content consumption. From a network analysis perspective, the set of interactions between members of these communities form *deviant networks* that map how the deviant content is shared and consumed. It is commonly believed that deviant networks are small and isolated from the mainstream social-media life; accordingly, most research studies have considered them in isolation.

We focus on adult content consumption networks, which is one *deviant network* with a significant presence in on-line social media and in the Web in general. We investigate two large on-line social networks and discuss the following insights. *Deviant networks* are limited in size, tightly connected and structured in subgroups. Nevertheless, content originated in *deviant networks* spreads widely across the whole social graph possibly touching a large number of *unintentionally exposed* users, such that the average local perception is that neighboring users share more deviant content. Finally, we investigate how content production and consumption varies with age and

Mauro Coletto
IMT Lucca - Ca' Foscari University of Venice
E-mail: mauro.coletto@imtlucca.it

Luca Maria Aiello
Nokia Bell Labs
E-mail: luca.aiello@nokia.com

Claudio Lucchese
CNR Pisa
E-mail: claudio.lucchese@isti.cnr.it

Fabrizio Silvestri
CNR Pisa
E-mail: fabrizio.silvestri@isti.cnr.it

show that the consumption rate is very similar between male and female users up to the age of 25.

We conclude that deviant communities are deeply rooted into the relational fabric of a social network, and that a deeper understanding of how their activity impacts on every other user is required.

Keywords Deviant network · Deviant behaviour · Pornography · Adult content consumption · Sexual content production · Social Media · Online social network · Tumblr · Flickr

1 Introduction

The structure of online social networks is fundamentally related to the interests of their members. People assort spontaneously based on the topics that are relevant to them, forming social groups that revolve around different subjects. This tendency has been observed with quantitative studies in several online social media [4, 47]. In the past, researchers have explored the relationship between information diffusion and network structure [8], focusing on the structural and dynamical properties of specific topical communities such as groups supporting political parties [20], or discussion groups about rumors, hoaxes [60] and conspiracy theories [9].

Online social media are also favorable ecosystems for the formation of topical communities centered on matters that are not commonly taken up by the general public because of the embarrassment, discomfort, or shock they may cause. Those are communities that depict or discuss what are usually referred to as *deviant behaviors* [18]; these are conducts that are commonly considered inappropriate because they are somehow violative of society's norms or moral standards. Pornography consumption, drug use, excessive drinking, eating disorders, or any self-harming or addictive practices are all examples of deviant behaviors. Many of them are represented, to different extents, on social media [22, 35, 52]. However, since all these topics touch upon different societal taboos, the common-sense assumption is that they are embodied either in niche, isolated social groups or in communities that might be quite numerous but whose activity runs separately from the mainstream social media life. In line with this belief, research has mostly considered those groups in isolation, focusing predominantly on the patterns of communications among community members [70] or, from a sociological perspective, on the motivations to that make people join such groups [7].

In reality, people who are involved in deviant practices are not segregated outcasts, but are part of the fabric of the global society. As such, they can be members of multiple communities and interact with very diverse people, possibly exposing their deviant behavior to the public. We aim to go beyond previous studies that looked at deviant groups in isolation by observing them *in context*. In particular, we want to shed light on three matters that are relevant to both network science and social sciences: *i)* how much deviant groups are structurally secluded from the rest of the social network, and what are the characteristics of their sub-groups who build ties with the external world; *ii)* how the content produced by a deviant community spreads and what is the entity of the diffusion which reaches users outside the boundaries of the

deviant community who voluntarily or inadvertently access the adult content; and *iii)* what is the demographic composition of producers and consumers of deviant content and what is the potential risk that young boys and girls are exposed to it.

In this initial study we undertake to answer those questions focusing on the behavior of *adult content* consumption. Public depiction of pornographic material is considered inappropriate in most cultures, yet the number of consumers is strikingly high [63]. Despite that, we are not aware of any study about the interface between adult content communities and the rest of the social network. The approach followed in this study looks at a deviant network *in context*. The same kind of analysis could be conducted on any other *deviant* topic.

We studied this phenomenon on two large dataset sampled from the Tumblr and Flickr social networks. The Tumblr dataset contains more than 130 million users and almost 7 billion directed dyadic interactions, while the Flickr dataset contains more than 39 million users and almost 600 million directed dyadic interactions. In both cases, we selected *deviant* users with a vocabulary-based approach. In Tumblr we used a large sample of 146 million queries from a 7-month log of search queries from Yahoo Search, from which we identified Tumblr pages clicked in response to *deviant* queries. In Flickr we followed a similar approach by looking at image tags. The adult dictionary is made publicly available [19].

Results show that:

- The deviant network is a tightly connected community structured in subgroups, but it is linked with the rest of the network with a very high number of ties (Section 4.1).
- The vastest amount of information originating in the deviant network is produced from a limited core of nodes but spreads widely across the whole social graph, potentially reaching a large audience of people who might see that type of content unwillingly, depending on the sharing actions enabled by each social platform. Although the consumption of deviant content remains a minority behavior, the average local perception of users is that neighboring nodes reblog more deviant content than they do (Section 4.2).
- There are clear differences in the age and gender distributions between producers and consumers of adult content. The differences we found are compatible with previous literature on adult material consumption: producers are older and more predominantly male and age greatly affects the consumption habit, strengthening it in males and weakening it in females (Section 4.3). Moreover if we look at age distribution of the consumers we recognize a similar consumption pattern both in Tumblr and in Flickr which is very different gender by gender: for male users the consumption of adult content increases with age until a maximum around 40-55; for female users the consumption increases only in youth to progressively substantially decrease. Relatively to the population of each social network analyzed we verify a higher consumption of pornography by male users, but interestingly male and female have similar consumption relative volumes up to the age of 25 on both networks.

Summary of the contributions. To the best of our knowledge, this is the first study that analyzes the production and consumption of adult content in general-purpose

online social networks, at very large scale. The computer science and social science literature in this area has focused so far on specific, small-scale deviant communities of nodes without analyzing this type of behavior in the broader social context these communities are immersed in. We study the phenomenon of adult content production and consumption for the first time *in context* by considering the whole social network of active users that surrounds the producers of pornographic material. The dataset we study is original, includes two social networks different in scope and structure, and it has no precedents for its size. This unique experimental setup allows us to study for the first time to what extent adult content spills over the restricted groups and spreads in the network; what we find provides strong evidence against the common assumption that adult content is mostly confined inside groups dedicated to that specific topic and provide a detailed analysis of the dynamics that characterize the spreading of adult content. Furthermore, the scale of the data allowed us to answer some longstanding social science questions around demographic patterns of pornographic material consumers: this is the first study that provides a very large scale quantitative measurement of adult content consumption across age and gender.

2 Related Work

Our contribution is related with previous work aimed at characterizing community dynamics, deviant content consumption, and potential impact of deviant communities on people's behavior.

Groups in online social media.
Computer science research has dealt extensively with the problem of classification of groups along structural, temporal, behavioral, and topical dimensions [2, 33, 53]. The relationship between group connectivity and shape of information cascades has also been explored, revealing an intertwinement between community boundaries and cascade reach that is particularly tight in communities built upon a common theme shared by all of their members [8, 27, 50, 61]. The degree of inter-community interaction has been analyzed mostly in the context of heavily polarized networks, the most classical example being online discussions between two opposing political views [1, 20, 29]. These studies explored methods to quantify segregation [34], but mainly focus on networks formed by two main divergent clusters.

Deviant communities.
A body of work has investigated the dynamics of misbehavior in online communities including newsgroups [57], question-answering portals [40], chats [67], and multiplayer video games [10, 66], trying to quantify the negative impact of misconducts on the community's health [21, 26, 72].

Studies about the depiction of drug and alcohol use in social media adopted mainly the content perspective. Researchers aimed at identifying the elements that boost content popularity, investigated the effect of gender on engagement, and studied the perceptions that deviant content arises in the young public [52]. Research has been conducted around anorexia-centered online communities [12, 32, 59], also

on Tumblr [22], investigating a wide range of aspects including the construction and management of member identities, the processes of social recognition, the emergence of group norms, and the use of linguistic style markers. Similar studies have been published over the years on communities of self-injurers and negative-enabling support groups, in which members encourage negative or harmful behaviors [35]. Fewer studies touch upon network-related aspects. One notable example is the work by Gareth et al. ([70]) that provides an overview of behavioral aspects of users in the PornHub social network, with particular focus on the role of sexuality and gender. More loosely related are studies on the so-called *dark networks*, mostly motivated by the need of finding effective methods to disrupt criminal or terroristic organizations [74]. The study by Christakis et al. ([17]) about the communication network between smokers and non-smokers is one of the few quantitative studies that addresses the interaction between the social network and one of its sub-groups, but it strongly focuses on the phenomenon of contagion.

Adult content consumption.
In the context of internet pornography consumption, computer science literature studied the categorization of content and frequency of use [38, 65, 69]. A wider corpus of research has been produced by social and behavioral scientists by means of surveys administered to relatively small groups. Special attention has been given to the relationship between age/gender and the exposure (voluntary or unwanted) to the internet pornography [13, 15, 51, 63, 75], with particular interest to the age range of young teens [15, 51, 73]. Numbers vary substantially between studies, but clearly men are more exposed than women (approximately 75%-95% vs. 30%-60%), with men exposed more frequently [36] and women more often involuntarily. It is estimated that young teens that are often exposed accidentally (roughly 25% to 66% of the times) and are also exposed to violent or degrading pornography (20% among female, 60% among male) [62]. Researchers have also pointed out the potential harm that adult material consumption through internet can cause, including addiction [42] and increased chance of adopting aggressive behavior [5]. Exposition also correlates with drug use [75] and with lack of egalitarian attitude towards the other sex [37]. Although delving into the potential harm of pornography is far beyond the scope of our work, this inherent risks provide an additional motivation to focus on this particular type of deviant community.

3 Deviant graph extraction

Online social media platforms provide users the capability of *publishing* content, accessing content published by other users of the network, and interacting with them. We model the social network as a graph whose nodes are the users and whose edges are the observed *interactions* among them. We distinguish among three kinds of interactions: *i) following* users enables a to receive updates from their content stream; *ii) liking* any piece of content produced by others expresses explicit interest in the activity of others; *iii) sharing* content produced by others increases the visibility of an item by re-posting it on the sharer's feed.

Each specific social network may re-brand these actions to fit the goals of the network; this study focuses on two online platforms: Tumblr and from Flickr. Tumblr is a popular micro-blogging platform where users can publish content by *posting* new entries on their blogs usually containing multimedia content, they can share content by *reblogging* any other post on their blogs, and *follow* other users. Flickr is a photo-sharing platform, including both amateur and professional users. Users can publish content by *posting* new photos or videos in their *stream*. They cannot re-share content but they can *like* photos and *follow* other users. Those platforms also provide some other actions, e.g., commenting a photo, but the analysis of rich textual signals goes beyond this work.

Tumblr and Flickr are ideal platforms for this type of study for three main reasons. First, they are two general-purpose sites that contain a wide variety of topical communities, they do not focus on few specific themes. Second, they do not enforce any content restriction policy around pornographic imagery (unlike Instagram, for example). Last, the two platforms are very different in scope, size, typical usage, and demographics of the user base, which helps us drawing more general conclusions.

We note that, in both platforms, the same human user could, in principle, own multiple user accounts (aka blogs). For the purpose of this study we consider blogs as users, and we will use the two terms interchangeably. We believe blogs are a good unit of analysis for the purpose of this study. All the dynamics of both social networks under investigation happen at the blog level: the following relationships and all other social actions (e.g., liking, reblogging) are done among blogs. Given that, it would be undesirable to coalesce different blogs owned by the same person into a single node. Furthermore, accurately matching blogs to actual users is hard even if one had the full information about all users. In fact, users can register their multiple blogs under different emails and using different nicknames. Historically, previous work on Flickr and Tumblr adopted the same assumption as we do and studied the network of blogs, not the network of users [22, 53].

We consider as *deviant nodes* those users who publish content about a given *deviant topic*: in our study the deviant behavior under analysis is the pornography production and consumption. To identify deviant nodes we resort to data from Web search engine logs and tags. We first discuss the methodology adopted on the Tumblr data, and then how this was adapted for the Flickr network.

As shown in previous studies analyzing pornography consumption in the context of Web search [45], the higher the number of deviant queries *hitting* (i.e., leading to the click of) a page, the higher the probability that page contains deviant content. In Tumblr, to identify a pool of candidate *deviant nodes* we consider deviant queries hitting the URLs of Tumblr blogs.

We used a seven-month long anonymized query log from Yahoo search engine, from which we collected a random sample of 146M US query log entries whose clicked URL belongs to the `tumblr.com` domain. After a simple query normalization process we obtained about 26M unique queries that hit a total of 2.7M unique Tumblr blogs. As expected, the distribution of number of queries hitting a blog is very skewed, with most popular blogs being reached by hundreds of thousands of clicks originating from search queries (Figure 1).

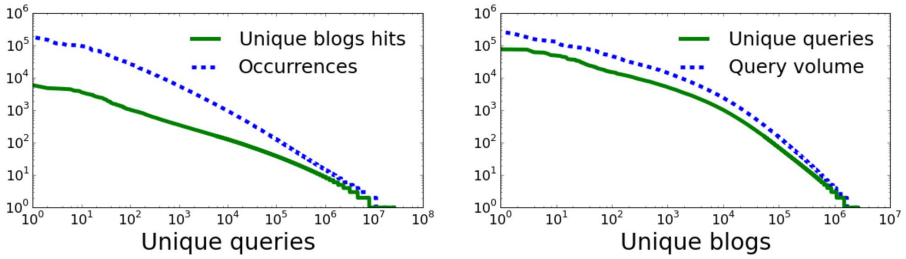

Fig. 1: Distributions of: (left) number of blogs hit by a query and number of occurrences of a query; (right) volume of (unique) queries hitting a blog.

To maximize the accuracy and coverage of the set of discovered deviant nodes, we devise an iterative semi-supervised *Deviant Graph Extraction* procedure [19]. The procedure detects adult queries by mean of a dictionary of terms: adult-marked queries are used to detect adult blogs which are their landing pages and the process is iterated until convergence extending the dictionary with new terms extracted by frequent queries which point to highly adult blogs previously detected, i.e. blogs whose incoming query volume is almost entirely pornographic. Figure 2 shows that the *Deviant Graph Extraction* procedure converges quickly. We were able to identify $198K$ adult blogs, through $4.2M$ unique queries filtered by mean of a dictionary which was expanded at each iteration of the procedure including in the final step $7{,}361$ terms.

Unlike in Tumblr, our Flickr datasets contains user-generated textual annotations that carry information about the content of published pictures, which allows us to directly estimate the pool of deviant content without the need of relying on external proxies[1].

To detect adult photos, we matched the vocabulary obtained through the *Deviant Graph Extraction* procedure with Flickr tags. Each picture in Flickr is labeled with manual tags. On average each picture has 6 tags and each user publishes around 122 pictures. We slightly modified the adult vocabulary to remove misleading words in a photographic context (e.g. black and white) and we filtered photos labeled with at least one tag in the adult dictionary. We considered only users with at least two public adult photos identified as above in line with the query approach used for Tumblr where we marked a blog as adult only if it was reached through at least two unique adult queries.

This procedure resulted in about 6.5 million photos by about $73K$ deviant users. In Flickr users share content not only in their profile but the platform enables the creation of groups whose member share images according to the topic of the group. To improve the recall of the data collection, we also identified those groups which *i)* include at least one of the previously detected adult users and *ii)* with a group title matches at least one word in our adult vocabulary. All the users and photos in such groups were included in the adult cluster. In so doing, about than 10M photos and $175K$ *deviant* users were detected in Flickr.

[1] Flickr users can mark their own photos as "adult". We first attempted to use this self-reported information to detect adult photos. We found that this approach leads to many false positives, mainly because very often pictures are marked in big batches containing adult and non-adult pictures.

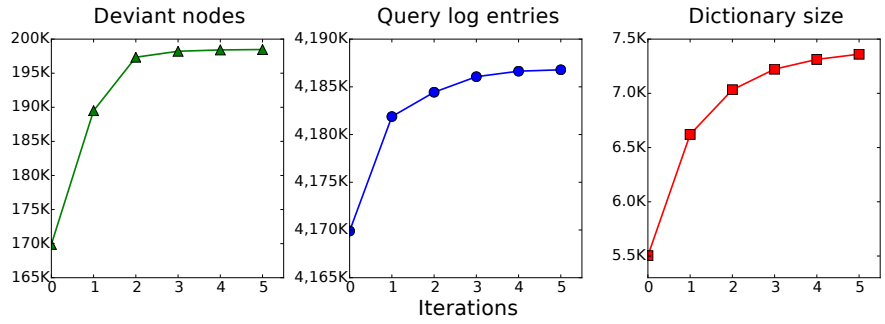

Fig. 2: Convergence of number of deviant nodes, query log entries and adult dictionary size during the Deviant Graph Extraction procedure.

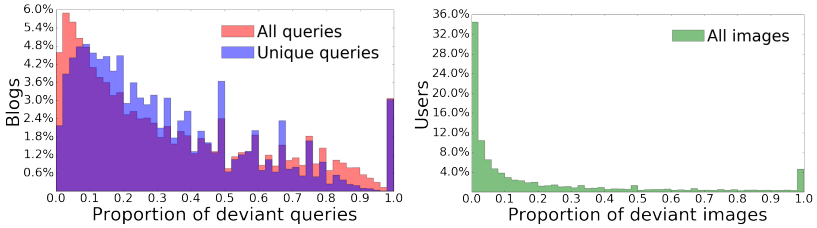

Fig. 3: (left) Distribution of deviant query volume ratio reaching deviant nodes in Tumblr, and (right) distribution of deviant photo volume ratio by adult user (publisher) in Flickr.

Table 1: Deviant users detected for Tumblr and Flickr:$|U|$. $|U^*|$ are users which are connected at least with one of the other users in $|U|$ through an incoming or outgoing reblog link (Tumblr) or favorite link (Flickr). $|U^*|$ in GCC are users in the giant component considering the reblog graph for Tumblr and the favorite graph for Flickr.

	$\|U\|$	$\|U^*\|$	$\|U^*\|$ in GCC
Tumblr	105K	75.6K (72%)	75.2K
Flickr	171K	156.1K (91%)	156.1K

In Figure 3 we report the distribution of the *deviant query volume* ratio for Tumblr and the *deviant image volume* ratio for the deviant users detected in Flickr. For both cases we see that the distribution is skewed, showing a tail of blog/user which are hit by a majority of deviant queries (Tumblr) or who published a high portion of adult pictures over their uploads.

To study the interaction of deviant nodes with the rest of the social network, we extracted a subset of the Tumblr and Flickr social network with a snowball expansion starting from the identified deviant nodes up to 3-hops away. Deviant nodes detected are reported in Table1. For Tumblr we considered the following and reblog actions while for Flickr since the reblog action is not present we collected information about favorites (or likes), which express a similar engagement even though they do not participate in the propagation of the content. The Tumblr follower network is a snapshot

Table 2: Network statistics for the reblog (R), follow (F), and favorite (L) networks of the full graph sample (*All*), the deviant graph (*Deviant*), and the communities that compose it (*Producers* and *Bridges*). All the statistics are about the giant weakly connected components and count only links whose both endpoints are in the considered node subset. $\langle k \rangle$=average degree, D=density, ρ=reciprocity, C=clustering, $\overline{spl}$=average shortest path length, d=diameter.

		$\lvert N \rvert$	$\lvert E \rvert$	$\langle k \rangle$	D	ρ	C	$\overline{spl}$	d
Tumblr	**All R**	14M	472M	33	$2{\cdot}10^{-6}$	0.06	-	-	-
	All F	130M	6,892M	53	$4{\cdot}10^{-7}$	0.10	-	-	-
	Deviant R	105K	1.4M	13	$1{\cdot}10^{-4}$	0.04	0.10	3.73	11
	Deviant F	135K	24.6M	182	$1{\cdot}10^{-3}$	0.07	0.13	2.80	8
	Prod$_1$ R	48K	914K	19	$4{\cdot}10^{-4}$	0.04	0.09	3.44	9
	Prod$_2$ R	16K	305K	19	$1{\cdot}10^{-3}$	0.05	0.13	3.19	8
	Bridge$_1$ R	9K	36K	4	$5{\cdot}10^{-4}$	0.04	0.08	4.18	13
	Bridge$_2$ R	3K	32K	11	$4{\cdot}10^{-4}$	0.06	0.21	3.32	10
Flickr	**All L**	15M	553M	37	$2{\cdot}10^{-6}$	0.06	-	-	-
	All F	39M	566M	15	$4{\cdot}10^{-7}$	0.26	-	-	-
	Deviant L	171K	13.4M	79	$5{\cdot}10^{-4}$	0.03	0.17	3.06	9
	Deviant F	169K	37.9M	224	$1{\cdot}10^{-3}$	0.28	0.21	2.77	9
	Bridge$_1$ L	66K	2.7M	47	$6{\cdot}10^{-4}$	0.05	0.17	3.05	13
	Prod$_1$ L	53K	4.6M	99	$2{\cdot}10^{-3}$	0.03	0.18	2.83	13
	Prod$_3$ L	20K	1.5M	94	$4{\cdot}10^{-3}$	0.03	0.23	2.53	13
	Prod$_2$ L	16K	1.0M	83	$4{\cdot}10^{-3}$	0.04	0.28	2.52	14

of the graph done in December 2015; the reblog network was built from the reblog activity happened in the same month. The Flickr follower network is a snapshot of the graph done in March 2016; the favorite network was built from the like activity happened until the snapshot time. The resulting networks are discussed in the Section 4.

We also obtained information about self-declared age and gender for about $1.7M$ Tumblr users and $12.3M$ Flickr users. The datasets include exclusively interactions between users who voluntarily opted-in for such studies. All the analysis we report next was performed in aggregate and on anonymized data.

4 Deviant graph in context

The availability of data about the interaction between deviant nodes and the social network that surrounds them provides the unique opportunity to study the structure and dynamics of a deviant network within its context. We first analyze the shape of the deviant network and measure its connectivity with the rest of the social graph (Section 4.1). We then look into how the information originating from deviant networks spreads across the boundaries of the deviant group (Section 4.2). Last, we study some demographic properties that characterize producers and consumers (Section 4.3).

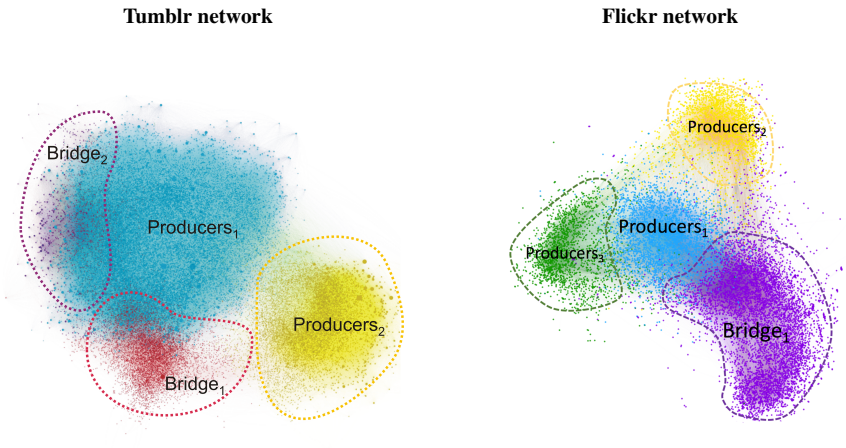

Fig. 4: Bird-eye view of the deviant network for Tumblr (reblog network, left) and Flickr (favorite network, right) with colors and labels denoting algorithmically-extracted communities.

4.1 Deviant network connectivity

The deviant network is a tiny portion of the whole graph, representing about 0.8% of all the nodes in the reblog graph in Tumblr, 1.1% of all the nodes in the favorite graph in Flickr and a even smaller portion in the follow network (0.1% in Tumblr and 0.4% in Flickr). So few nodes could be scattered along the social network or clustered together. So we ask:

Q1) Are deviant nodes organized in a community?
We consider the *deviant networks* as the subgraphs of the follow and reblog/favorite networks induced by the *deviant nodes* in Tumblr and in Flickr. A directional link in the follow (reblog/favorite) network from node i to node j exists if i follows (or reblogs/likes the posts of) j, meaning that the information flows from j to i. Basic network statistics on such subgraphs reveal that the deviant networks are quite dense, yet they have a high diameter (Table 2). Similar statistics have been observed before in other social networks [3] and might be an indication of the presence of strong sub-groups patterns, as well as a signal of the absence of a community structure. To better determine the reason for such elongated shape, we run the Louvain community detection algorithm [11] on the deviant network[2]. We considered the reblog network for Tumblr and the favorite network for Flickr as they are more representative than follow network of the dynamic interactions of the users in the OSN. Only the giant connected component of the network has been considered which corresponds to 72% of the network in Tumblr and 91% in Flickr (see Table1). The modularity measure for the clustering in Tumblr is 0.44 and 0.53 in Flickr; those are high values of modu-

[2] Louvain is a modularity-based graph clustering algorithm that shows very good performance across several benchmarks [31] and that is fast to compute even on large networks.

larity [54] that indicate the presence of structured, well-separated communities. Four clusters emerge both in Tumblr and in Flickr, whose network statistics are summarized in the bottom lines of Table 2. To determine their nature, we manually inspected the content of 250 blogs in each of them.

In Tumblr more than 90% of all the blogs in the two largest clusters contain blogs that *exclusively* produce explicit adult content, aimed at an heterosexual public (*Producers*$_1$) or at a male homosexual public (*Producers*$_2$). The blogs in the two remaining communities post less explicit adult content and more sporadically, often by means of reblogging. They either focus on celebrities (*Bridge*$_1$), or function as aggregator blogs with high content variety, including depiction of nudity (*Bridge*$_2$). In Flickr, the composition is different with similar attributes. We found (*Producers*$_1$) and (*Producers*$_2$) clusters with the same content characterization described for Tumblr and in addition to them a new cluster has been identified (*Producers*$_3$) whose users share mainly pictures representing transvestites or transsexuals. The same cluster is likely present in Tumblr but its size is not large enough to be distinguished by other producers. Producers clusters in Flickr are less than 58% of the deviant nodes with a large bridge cluster (*Bridge*$_1$) which is characterized by a content less explicit (soft porn, artistic nudity, manga).

From a bidimensional visualization of the network layout (Figure 4) it becomes apparent that the *Producers*$_1$ and *Producers*$_2$ are two well-separated cores. In Flickr *Producers*$_3$ is very close to *Producers*$_1$. The remaining communities are peripheral and arranged in a crown-like fashion in Tumblr (which explains their high diameter) around the largest sub-cluster *Producers*$_1$; in Flickr instead the largest cluster is very elongated (showing the highest diameter) indicating a strong presence of soft content which from a network point of view is organized in long cluster connected with the hard part only with one side. We named all the non-producers groups *bridge communities* as their main focus is often not on deviant content and, as we shall see next, they act also as link towards the rest of the graph.

In short, we find that deviant nodes are not scattered in the social network but are tightly organized in a structure of distinct communities. To find out about the nature of their interaction with the rest of the social ecosystem, we proceed to answer the next question.

Q2) To what extent is the deviant graph connected to the rest of the social network?
There are several ways to estimate the connectivity between two sets of nodes in a graph. We use different metrics to measure it between the four communities of the deviant network and the rest of Tumblr, as summarized by the matrices[3] in Table 3; rows represent the group of nodes from which the social tie originates, columns those on which it lands.

The average volume of connections (left matrix) provides a first indication about the difference in connectivity across different groups. For both Tumblr and Flickr, the diagonal has the highest values because of the community structure of the deviant network and of its sub-communities: members of a group have many more ties to-

[3] Link directionality is considered: ties originate from groups listed on the rows and land on groups listed on the columns.

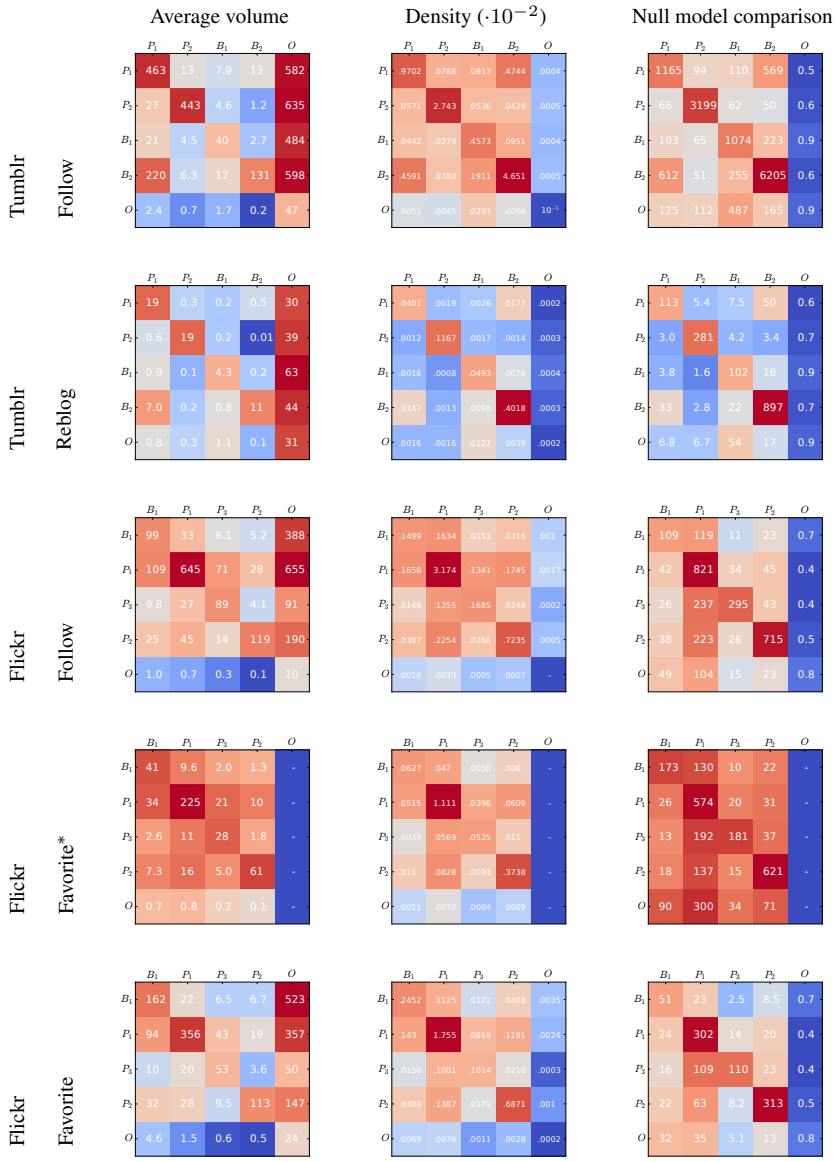

Table 3: (T=Tumblr, F=Flickr) Measures of connectivity between the communities in the deviant network in Tumblr (*Producers* P_1, P_2, P_3 (F) and *Bridges* B_1, B_2 (T)) and the rest of the social network O, for the follow, reblog (T), favorite to adult content or favorite* (F) and total favorite (F) relations.

wards other group members rather than to the outside. This is true in particular for $Producer$ clusters. In Tumblr the difference between the diagonal values and the other cells is more prominent, however the volume of links incoming to the largest producer cluster is particularly high from the smallest bridge community ($Bridge_2$), which surrounds it. The average Tumblr user in our sample (see rows O in Table 3) follows around 51 users, between 2 or 3 of which are in the core of the deviant network and around 2 of them are in bridge communities; similarly, among the 33 users reblogged in one month by the average user, one is from a *Producer* cluster and one from a *Bridge* group. In Flickr, also the values not in the diagonal are generally high showing an significant connection among clusters: a user follows on average around 12 other users and among them 1 is from the bridge cluster which is the biggest in size and 1 is a member of a producer cluster; a user likes on average pictures from 31 users and around 5 of them are in the bridge cluster and between 2 or 3 are among producers. Since the classification of deviant content in Flickr has been done at picture level through tags, we can distinguish likes to adult content (forth row in Table 3: favorite*) and total likes (fifth row). Among the users liked on average among the bridge cluster only 15% are favorite links on adult content, while for the producers clusters the percentage rises to 42% confirming that the content of the bridge cluster is less explicit and often not adult.

When looking at raw volumes, the amount of links from the deviant network to the rest of the graph is very high, mainly due to the high dimensionality of the set of nodes that are not deviant. To partially account for dimensionality of the groups, we measure the connectivity with density computed as the ratio of edges between the two groups over the total number of possible edges between them (Table 3, center). Also in this case the overall patterns hold, but the connectivity towards the external graph drops significantly.

Values of density are still affected by size, though. It is known that in real networks there is a strong correlation between number of nodes and graph density [48]. To fix that, in the spirit of established work in complex systems [64] we resort to a comparison of the real network connectivity with a *null model* that randomly rewires the links while keeping the degree of each node unchanged. The values we report in Table 3 (right) indicate the ratio between the real observed connections and the null model. Also in this case, values on the diagonal are very high (except for the outer network, which has a value close to 1, as expected). For Tumblr, both the density matrix and the null model comparison show an high connectivity not only for the producers (in particular $Producer_2$), but also for bridge communities. This is particularly evident for $Bridge_2$ both in the reblog and in the follow matrix. For Flickr instead the bridge community is characterized by the lower value both in favorite and follow in the diagonal among deviant clusters. Moreover if we look at $Producer_3$ in Flickr we discover that is highly connected to $Producer_1$: members of $Producer_3$ like and follow significantly members in $Producer_1$ cluster while the contrary is much less evident. This confirms the assumption that a similar cluster $Producer_3$ might be present for Tumblr but not identifiable since it is small or highly connected with $Producer_1$ resulting indistinguishable from it. Also, this computation highlights that ordinary users have a tendency in Tumblr to reblog content from the core of the deviant network almost 7 times more than random and between 17 and 55 times more

than random from the bridge community members, in Flickr to like content from $Producer_1$ almost 35 times more than random (300 times if we look at adult content only) and 32 times more than random from the bridge community members (90 times if we look at adult content only).

Last, we complement the connectivity analysis with a measurement of the betweenness centrality of nodes in the different communities of adult content producers (Table 4). The betweenness centrality [71] quantifies the number of times a node acts as a bridge along the shortest path between any two other nodes in the whole network, so it is a strong indication of the brokerage potential of a node with respect to information flow. In both networks the largest betweenness is observed for the bridge clusters, with a peak for the *Bridge*$_1$ cluster in Tumblr (celebrities). This further finding confirms that bridge cluster play an important role in diffusing adult content to the rest of the network.

Table 4: Average betweeness centrality C_b of nodes belonging to different adult communities in Tumblr's reblog network (R) and Flickr's like network (L). Centrality values are max-min normalized.

Tumblr	C_b
Prod$_1$ R	0.0038
Prod$_2$ R	0.0061
Bridge$_1$ R	0.0133
Bridge$_2$ R	0.0061
Flickr	C_b
Bridge$_1$ L	0.0074
Prod$_1$ L	0.0060
Prod$_3$ L	0.0053
Prod$_2$ L	0.0049

In summary, the core of the deviant community is dense but it is far from being separated from the rest of the graph, which is connected to it both directly and even more tightly through bridge groups.

4.2 Deviant content reach

We found that, although the deviant network forms a tightly connected community, it is not isolated from the rest of the social graph. This calls for an investigation about the visibility that the deviant content has in the outer network and what are the main factors that determine its exposure. We do so by answering four research questions.

Q3) How much deviant content spreads in the social graph and what are the main agents of diffusion?

The exposure to deviant content goes beyond the members of the deviant network who are the *producers* of original adult material. Specifically, the *consumers* of deviant content can be categorized in three classes. The first is the class of *active consumers*: nodes who reblog (but not necessarily follow) adult posts, thus contributing

to its spreading along social ties. Posts can be re-blogged in chains and create diffusion trees that potentially spread many hops away from the original content producer, therefore active consumers could further be partitioned in those who spread the content *directly* from the producers and those who do it from *indirect* reblogs. This class is present in Tumblr; in Flickr sharing actions are not made available but users can see the pictures liked by other people they follow in their feed. For this reason we called *active consumers* in Flickr all the users who like adult content. The favorite action is stronger than the following action which characterizes passive consumers. This influences the spreading dynamic which is limited in Flickr compared to Tumblr also for these infrastructure and visibility limitations. The second is the class of *passive consumers*: nodes who do not contribute to the information diffusion process but are explicitly interested in adult content because they directly follow the producer nodes or the like their content. In Flickr we will distinguish passive consumers by looking at the actions: follow or both follow and favorite actions to the adult content generated by producers. The last class is the one of *involuntary consumers* or *unintentionally exposed* users. In Tumblr, users in this class are the ones who do not follow any producer node and do not reblog their content, but happen to follow at least one active consumer who pushes adult content in their feed through reblogging. In Flickr, we consider unintentionally exposed all the users who follow people who liked adult content at least once; again this choice is motivated by the fact that the Flickr feed shows the pictures that neighbors recently liked. This way of estimating involuntary consumers users is a best-effort approach that provides an upper bound on the number of people who are actually exposed to adult content; as we do not have access to click logs, there is no way to know which pages users visited.

By drawing a quantitative description of the volume of deviant content reaching these three classes we can estimate how much the adult community is visible in the network at large. We adopt a conservative approach in which we consider the *Producers* communities as the only ones generating original explicit content. Given the results of the aforementioned manual inspection, we are very confident that their activity is mainly focused on the production of adult material.

Tumblr. We measure the size of the different consumer classes and the amount of content that flows through or to them by means of reblogging. The results are summarized by the schema in Figure 5. The network of deviant content producers is very small but receives a considerable amount of attention from direct observers. The audience of passive consumers counts almost 24M people. Around 2M users reblog directly from the deviant network, for a total of around 28M reblog actions in one month. A consistent part of the two *Bridge* communities within the deviant graph (a total of 3K users) are also direct consumers, and they reblog *Producers* 56K times per month. When looking at the set of 2.4M users who indirectly reblog deviant content, we see that only a small fraction of their monthly reblogs (less than 7%) is performed through bridge communities. However, in relative terms, bridge communities are considerably more efficient in spreading information than the average active consumer. If we consider efficiency η of a user set U as the ratio between reblogs done r_d and reblogs received r_r, weighted by the cardinality of the set $\left(\eta = \frac{r_r}{r_d \cdot |U|}\right)$, we discover that the bridge communities ($\eta = 1.5 \cdot 10^{-3}$) are several orders of magnitude more

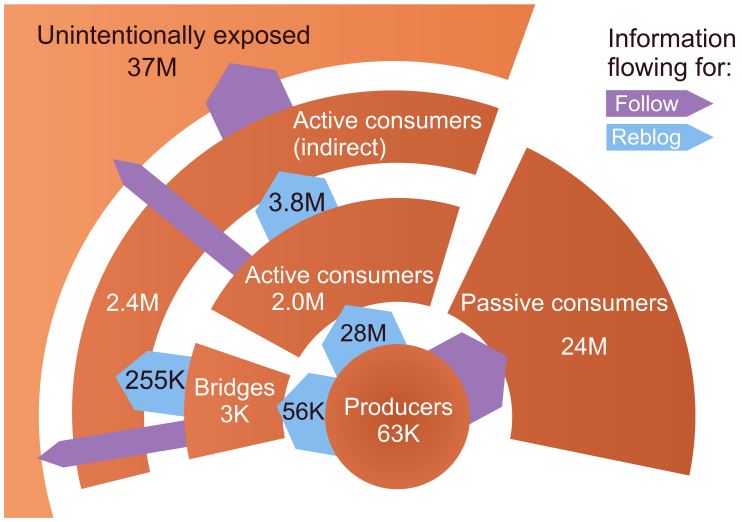

Fig. 5: Diffusion of deviant content from the core of Producers to the rest of the network in Tumblr. Sectors represent disjoint user classes and arrows encode the information flow between them. Reblog arrows report the total volume of reblogs between two classes.

effective in spreading the content farther away in the network than the rest of active consumers ($\eta = 6.7 \cdot 10^{-8}$). Last, the audience of users who are potentially exposed in an unintentional way to deviant content includes almost 40M people. This figure should be considered as an upper bound on the number of people who actually have been exposed, as a follower of an active consumer might not see the pieces of deviant content for a number of reasons (e.g., inactivity, amount of content in the feed). That said, the pool of people who are potentially exposed is still very wide.

Flickr. Similarly for Flickr we quantified the sizes of different classes. Because of the absence of resharing actions in Flickr we focused on *passive consumers* and *unintentionally exposed users* only. The results are summarized by the schema in Figure 6. The size of the producer clusters (90K) is almost 43% bigger than the case of Tumblr but still very small compared to the whole network which is composed by 39M users. The size of the passive consumers instead is smaller in comparison with the same class in Tumblr: around 2M users, accessing deviant content by following the producers. The favorite action is quite limited: only 226K users exclusively like producers' content and 475 K like and follow producers at the same time. For Flickr the bridge cluster has a role different from the bridges communities in Tumblr: resharing actions are not enabled but they have an important role since their content is an entry point to access adult content by navigating the social network and the followers. In particular around 37% of them follow the producers and around 48% of them are in group of users who like and follow the producers. Last, the audience of users who are potentially exposed in an unintentional way to deviant content includes almost

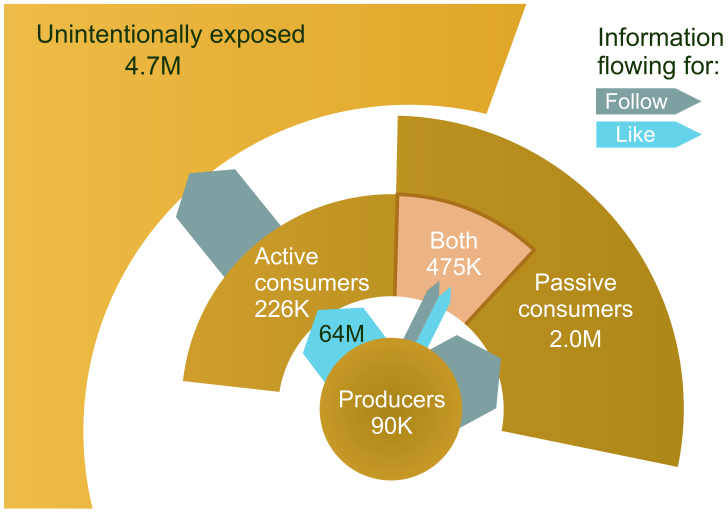

Fig. 6: Diffusion of deviant content from the core of Producers to the rest of the network in Flickr. Sectors represent disjoint user classes and arrows encode the information flow between them.

4.7M people. The members of this group like pictures from users who liked at least one picture shared by the deviant producers. The pool of people who are exposed to adult content in Flickr is significant but smaller than Tumblr's, in proportion, mainly because the platform enable less sharing tools (no reblog actions). Also, the two platforms have very different content-production dynamics. Manual inspection suggests that Tumblr deviant blogs tend to be more topically focused and contain multimedia material taken from other Web sources. On the contrary, deviant Flickr users are mostly amateur who publish pictures that they take, which makes their content more appealing to their social circles rather than to the general audience, thus restraining wide diffusion.

Content reach and node coreness. Past literature has shown how efficient information spreaders do not necessarily correspond to the most highly connected nodes in the network [14]. Kitsak et al. [41] have found instead that the top spreaders are often those located within the core of the network as identified by a *k-core* decomposition [71]. In network analysis, a *k-core* is a subgraph containing nodes of degree k or more *within* that subgraph. A *k-shell* is a set of nodes that belongs to the *k-core* but not to the $k+1$ core. A node that belongs to a given *k-shell* is said to have a *shell index* equal to k. The shell index is a way to measure the *coreness* of a node in the graph.

We compute the *shell index* k_s on the follower networks for every adult node and plot it against the total number of (direct and indirect) reblogs the node's content has received in Tumblr and against the total number of likes received in Flickr (Figure 7). In Flickr, we have found a trend that is similar to what has been found in previous

work, with the node reach growing as the shell increases, with diminishing returns. In Tumblr, interestingly, the reach is more spread across several orders of magnitude for fixed values of shell index. This is mainly given by the behavior of bridge clusters (B1 and B2). Bridge nodes have a higher reach also at lower values of shell index, which confirms their role as brokers in the process of content spreading: bridges take advantage of their high-betweenness rather than their high coreness to act as effective spreaders.

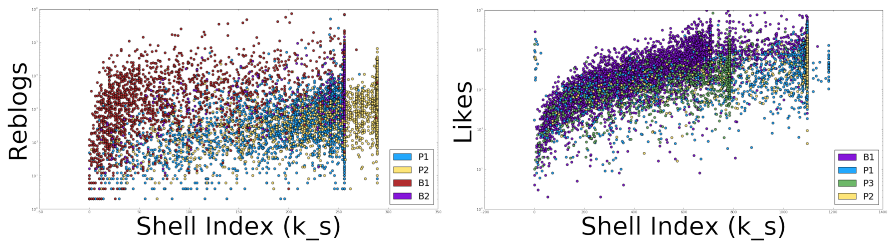

Fig. 7: Correlation between shell index k_s of the social graph (follower network) with the reblogs for Tumblr (left) and with the likes for Flickr (right). Each point represents a blog that produces adult content. Colors encode the different adult sub-communities.

Q4) What is the perception of deviant content consumption from the perspective of individual nodes?

Similar to real life, individuals in online social networks are most often aware of the activities of their direct social connections only but lack a global knowledge of the behavior of the rest of the population. In fact, the broad degree distribution of social networks may lead to the over-representation of rather rare nodal features when they observed in the local context of an ego-network. This phenomenon has been observed in the form of the so-called *friendship paradox* [28, 39], a statistical property of social networks for which on average people have fewer friends than their own friends. More recently the concept has been extended by the so-called *majority illusion* [46], which states that in a social network with binary node attributes there might be a systematic local perception that the majority of people (50% or more) possess that attribute even when it is globally rare. As an illustrative example, in a network where people drinking alcohol are a small minority, the local perception of most nodes can be that the majority of people are drinkers just because drinkers happen to be connected with many more neighbors than the average. In our case study, active deviant content consumption is definitely a minority behavior compared to millions users of our sample both in Tumblr and Flickr.

To estimate the presence of any skew in the local perception of deviant content consumption, we consider the nodes who are not producers and calculate the distribution of the proportion of their neighbors (in both the follow and reblog graphs) that either produce or reblog deviant material. The result is summarized in Figure 8.

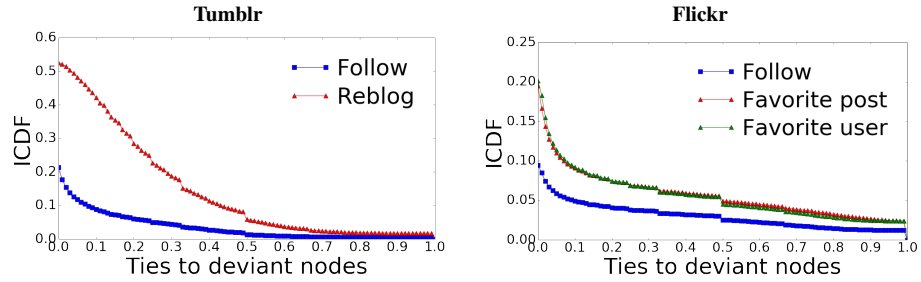

Fig. 8: Proportion of nodes with at least a given ratio of outlinks landing on deviant nodes (inverse cumulative density function) in Tumblr (left) and Flickr (right).

We observe that the follower network is nowhere close to exhibit the majority illusion phenomenon, with only the 10% of the population having 10% or more of their neighbors posting or reblogging deviant content in Tumblr and with only the 5% of the population having 10% or more of their neighbors posting or liking deviant content in Flickr. The effect increases sensibly when considering the reblog network in Tumblr, with 40% of the population locally observing more than 10% of their contacts reblogging deviant content and almost 10% having more than half of their neighbors doing it. This happens partly because the size of the reblog network is one order of magnitude smaller than the one of the follower network, as we consider reblogging activity for one month only. In Flickr the majority illusion phenomenon is less prominent even though if we look at the favorite graph the effect is doubled compared to the follow graph; this means that when looking at recent activity only (reblogs or likes), local perception biases are much stronger (although not predominant) in the community than what can be inferred from the static follow graph.

Although strongly biased perceptions are not predominant when counting the number of neighbors, a stronger bias emerges when looking at the *volume* of deviant content that is observed by a node from its neighbors in Tumblr. We calculated that more than 71% of nodes in Tumblr reblog less deviant content than the average of their friends (considering friends who posted or reblogged at least once in the time frame we consider). This effect, that derives directly from the strong correlation between degree and number of posts and reblogs, suggests that the local users' perception of other people's behavior is skewed towards an image of pervasive consumption of deviant content in Tumblr that might be a driver to stimulate the diffusion of deviant content in this social network.

Q5) How cascades of deviant content are characterized?

We have found that the activity of a relatively small groups of producers can echo in the network and reach a very large audience. An open question remains about how an individual piece of deviant content spreads along the social ties in comparison to any other content type. To partly answer this question we focused on Tumblr, where a post can be reblogged by several users, possibly in long chains, thus generating *information cascades* [25, 49].

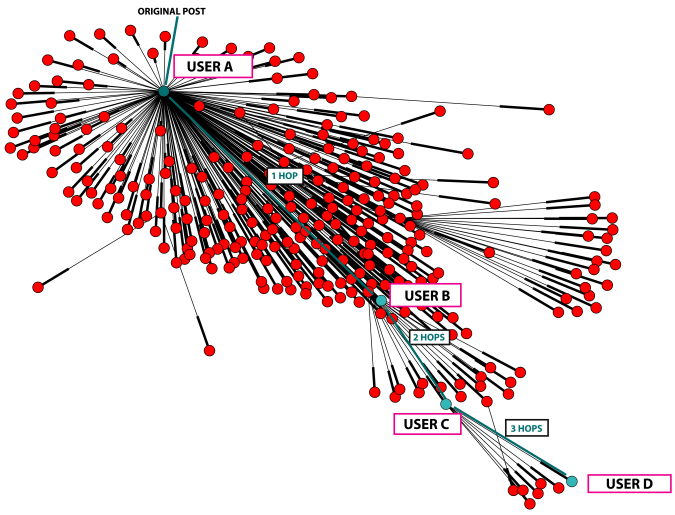

Fig. 9: Example of a reblog cascade generated by a post published by user A. In particular a chain is highlighted in light blue showing user D reblogging the post generated by user A through user C, who had reblogged the content from user B before. The chain is three-hops long.

Figure 9 show an example of a cascade generated by a post published by user A. All the nodes represent other users sharing that post directly from the original blog or through other users who previously had shared the content (e.g., user C from user B). The nodes without outgoing links are the leaves of the tree, final re-blogging actions whose content is not reshared anymore in the period of analysis. In the picture a chain of 3 hops is highlighted. We are interested in adult posts which extensively propagate, with long reblog chains (> 10 hops).

We selected 157K posts created in the first week of January 2016 by users in any of the *Producer* clusters and that are reblogged at least once. Figure 10 shows the cascade size and depth distribution for a sample of the selected adult posts and for non-adult posts. The distributions are in line with similar results obtained in previous studied related to cascades [16] with a weaker propagation effect for adult content cascades compared to regular type of content.

Most posts are reblogged just for a few days after they are published, even though there is a consistent tail of posts that gets reblogs for several weeks (Figure 11). We focus on the 529 posts that generate long cascades with 10 hops or more. We compare those with 657 non-adult posts (manually checked) with comparable virality (10 hops or more) published by users in the *Bridge* cluster.

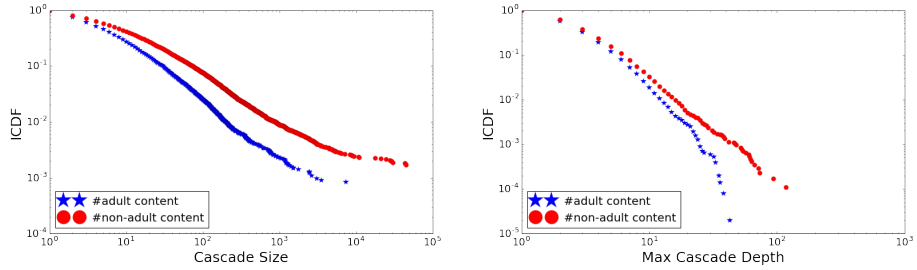

Fig. 10: The inverse cumulative distribution function (ICDF) of cascade size and depth for a sample of adult and non-adult posts.

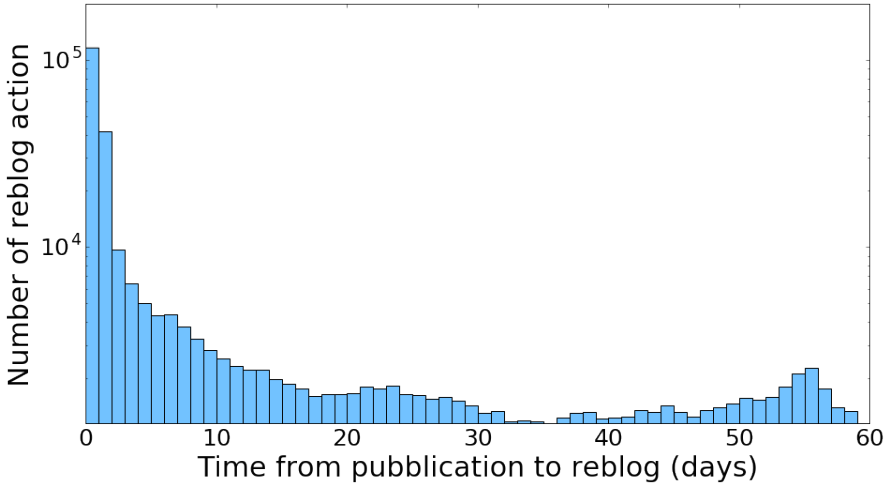

Fig. 11: Number of reblogs of adult posts created by Producers in the first week of January 2016. We considered all the reblog actions happening in January and February 2016 regarding the selected 157K posts created in the first week of January 2016. For each reblog action we plot the time in days from the publication date of the original root post to the reblog action time. The histogram shows the number of reblogs (y-axis) in relation to the amount of days passed by the original publication of the reblogged content.

We characterize each cascade by the maximum time of diffusion (time of the last reblog) and the number of leaves that the post generated (number of final users reblogging the content which is not reshared anymore from them). Figure 12 shows the relationship between the two dimensions, for maximum time of diffusion binned in 10-days long intervals.

The time of last reblog is an indicator of the persistence of a content item over time. We observe that the persistence in time is similar both for adult and non-adult content. We use the number of leaves to understand how the frontier of the cascade evolves over time. In this case, the frontier of adult content is much smaller that non-

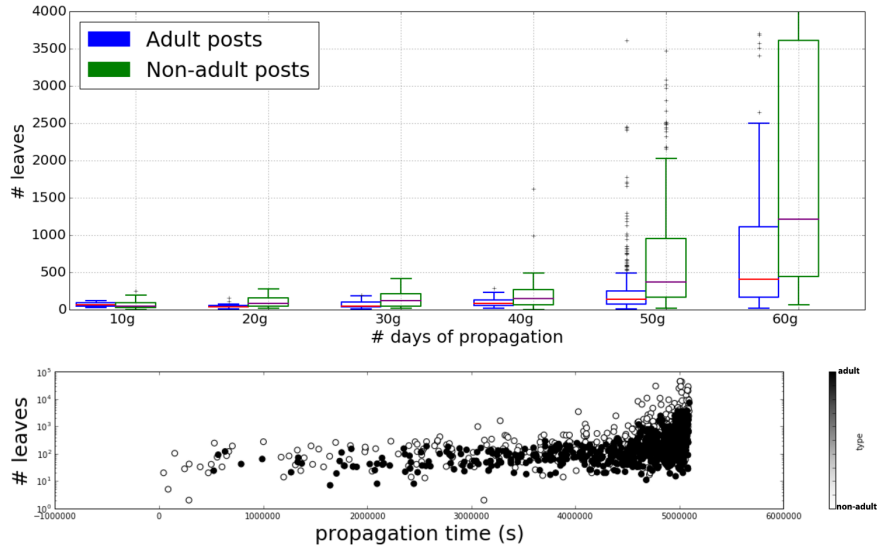

Fig. 12: Time length vs number of leaves for porn and not porn post cascades (> 10). Each boxplot (blue for adult posts and green for non-adult posts) shows how all the reblogs done in the temporal range (days from publication of the original post and the reblog action) are distributed (red line is the median and 50% of the data are contained in the box).

adult, meaning that adult content is less likely to be reblogged ant that the cascade as a smaller width. This might indicate that adult content can become viral and it reaches users far in time and number of hops but the process involves less people than for non-adult content probably because of the embarrassment of sharing this kind of content which shows that drivers of the information are less but still they spread a lot the content. We verified statistically that the difference in number of leaves given a time diffusion for a post taken from the two samples (adult and non-adult) is significant through the Welch Two Sample t-test done on the ration leaves over time. The test is significant with a low p-value ($p = 867e - 07$).

Q6) Is it possible to reduce the diffusion of deviant content in Tumblr with targeted interventions?

Previous literature that investigated the properties of small-world networks indicates that information spreading or other phenomena of contagious nature can be drastically reduced by acting on a limited number of nodes in the graph [56]. Effectiveness of targeted interventions has been shown in a variety of domains, epidemics being the most prominent among them.

The intuition informed by previous work suggests that the wide diffusion of deviant content can be reduced by properly marking the posts produced by a small set of core nodes and showing them only to people who explicitly declared their interest for that specific topic. In a simplified experimental scenario, we measure the proportion of active consumers reached by adult content in a setting where all the posts from a

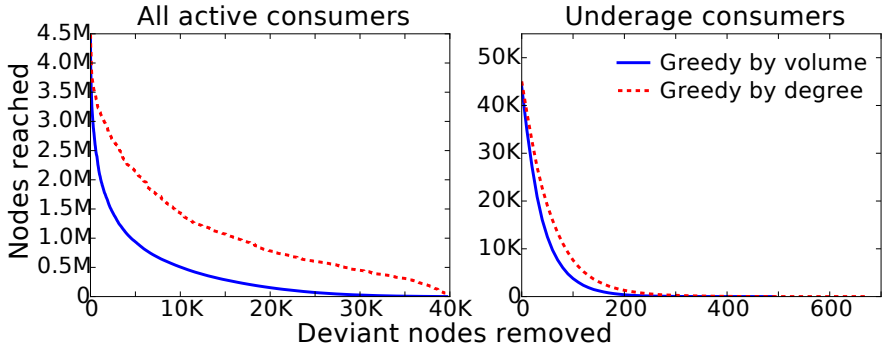

Fig. 13: Shrinkage of content diffusion after deviant nodes removal, using two different strategies.

set of core nodes C are erased. The question is how to select C and how big it needs to be to uproot the diffusion process.

The optimal selection of nodes is a set cover problem (NP-complete), but we test two common approximated strategies to solve it: *i) greedy by volume*, an algorithm that ranks nodes by the number of blogs that are reached by the content they produce; and *ii) greedy by degree*, that takes into account the network structure only and ranks nodes by their in-degree in the reblog network. The effectiveness of the two approaches as $|C|$ increases in shown in Figure 13. Although using the indegree as proxy for the diffusion potential is not optimal, the removal of the $5{,}000$ highest indegree nodes curbs the diffusion by more than 50%. As expected, the strategy by volume is more effective (as it better approximates the optimal set cover), with a surprisingly sharp decay of the deviant content reach. The removal of the $5{,}000$ top nodes reduces the information spreading by nearly 80%, which increases to almost 100% when extending the block to $25{,}000$ nodes. Furthermore, using our sample of demographic information, we find that to limit the exposure of underage users would be sufficient to remove the 200 top nodes, as identified by any of the two selection strategies. To monitor and control the capabilities a vertex may have on data flowing in the network other alternative approaches based on Routing Betweenness Centrality (RBC) have been proposed [23], but we limited to simple strategies since the intervention strategies to limit the flow of this content are not the main scope of this work.

We used Tumblr as a case study to show the effect of targeted interventions because in Flickr the spreading of the content is limited by the platform and the access of adult content is direct and not through cascades.

4.3 Demographics factors

The demographic composition of online adult content consumers has been measured by several sociological surveys (see Section 2), but none of them partitions the participants according to their type of consumption. Yet, we have shown that the categories

Table 5: Gender distribution among different categories in Tumblr and in Flickr.

	Group	Male	Female
Tumblr	All	29%	71%
	Producers	84%	16%
	Bridges	47%	53%
	Active consumers (reblog)	31%	69%
	Passive consumers (follow)	34%	66%
	Unintentionally exposed	20%	80%
Flickr	All	59%	41%
	Producers	76%	24%
	Bridges	72%	28%
	Active consumers (like)	87%	13%
	Passive consumers (follow)	68%	32%
	Unintentionally exposed	62%	38%

of people exposed to online deviant content range from the active content producers to unintentional consumers. This calls for an investigation of the relationship between type of consumption and demographic characterization.

Q7) Is there a significant difference in the distribution of age and gender between members of the deviant network and people with different levels of exposure to deviant content?

We report the distribution of age and gender of users with different levels of exposure to adult content, computed on the sample of 1.7M Tumblr users and 12.3M Flickr users who self-reported their demographic information. The two social networks are very different in the age distribution of the users in addition to typology of content. Tumblr is a social network targeted on young people. The average age in the sample is slightly higher than 26, and female are the majority (71%). Flickr is more used by adult people and professional photographers with an age on average of 41 and it is more balanced in the gender of the users with 59% of male users (Table 5).

To partly validate the user-provided information, we first compare them with third-party statistics. Our numbers are roughly compliant with several public reports that rely on orthogonal methods for assessing the age and gender of users (e.g., surveys and clickstream monitoring [44, 58]). Also, we further validate the gender data by assessing that the 95% of users in Tumblr and 95% of users in Flickr in the $Producer_2$ cluster focused on male homosexual content are indeed male. The overall age distribution of age by gender is shown in Figure 14 for Tumblr and for Flickr: in both OSNs male tend to be older with an higher age variance in Flickr. In Tumblr moreover the percentage on underage users is high: around 10% while in Flickr is 0.2%, this confirm why is important to study strategies to prevent visibility of adult content to minors as we have discussed in the previous research question especially for OSN popular among young people. Despite the spikes corresponding to birthdays in round decades (1970, 1980, and 1990), probably due to misreporting, the distributions still tend to be Gaussian, as expected.

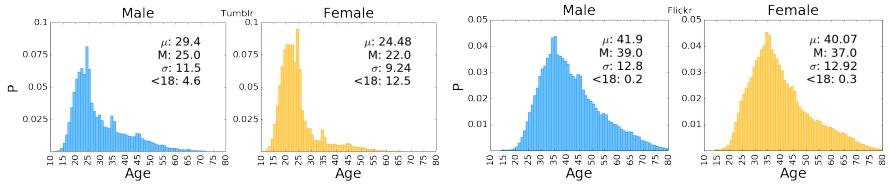

Fig. 14: Age distribution of users in our dataset for Tumblr (left) and Flickr (right). Mean μ, median M, standard deviation σ, and percentage of users under 18 years old are reported.

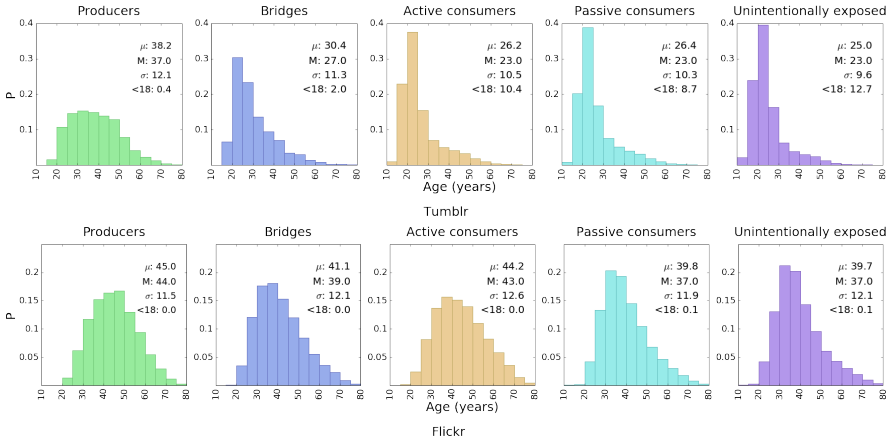

Fig. 15: Age distribution of different groups of producers and consumers of adult content in Tumblr (top) and in Flickr (bottom).

We then measure differences in age[4] and gender distribution for the user classes of *producers*, *bridges*, *active consumers*, *passive consumers*, and *unintentionally exposed* users (Figure 15). In Flickr we have the same categories with a different characterization as reported in Figure 6: Likers (active consumers), Followers (passive consumers).

Tumblr. Figure 15 shows that producers are considerably older than the typical user, averaging around age 38 and with almost no underage users. Different from the overall distribution, they are mostly male (84%), in alignment with studies indicating that men are more involved in assiduous consumption of adult material (Table 5). Moreover we found that bridge groups are fairly gender-balanced (with more female –53%– in the celebrity-oriented community) and include younger people (30 years old on average). Consumers of deviant nodes who actively reblog or passively follow deviant blogs are covered by demographic data at 12%, proportion that drops to 4% among those who follow deviant nodes. In both classes, the age is quite representative of the overall Tumblr population in our sample (more than 66% female). A similar

[4] The number of samples in each age distribution is high; therefore, as expected, all the differences between the average values are statistically significant ($p < 0.01$) under the Mann-Whitney test.

male-female proportion holds for people that are potentially exposed to deviant content in an unintentional way. This last class has the highest proportion of underage people (13%), which reinforces the concern about young teens unwillingly seeing inappropriate content.

Flickr. Figure 15 shows that producers are older than other categories (4 years more than average) and unintentional exposed users are the youngest as it is in Tumblr. Moreover we found that producers in Flickr are for the 76% male, confirming a higher presence of men in production of adult content. Among consumers we have differences in the average age in we consider people who like adult content (avg. age 44) and users who follow adult content producers (avg. age 40). This shows a different behavior in consumers: younger people tend to follow more than liking and male users tend to like more than female. Indeed, if we look at gender we surprisingly see that among likers 87% are male, while among followers only 68% (Table 5). The bridge community has a lower percentage of male users compared to the producers (72%) probably showing indicating a more tendency to consume hard content by male and soft content by female.

The fact that the gender distribution for consumers deviates only slightly from the overall gender distribution is in partial disagreement with previous studies on gender and sexual behavior [36, 43] which state that men are usually more exposed than women to adult material. This is particular evident in Tumblr and we conjecture that this might happen because of the tendency of female to have their peak of adult content consumption in a much younger age than men (as shown by [30]), combined with the predominance of young female among Tumblr users. To verify better the relationship between age and gende in consumption of adult content in both Flickr and Tumblr we aim to answer one more question.

Q8) Does age have an effect on how different genders consume adult content?
To find out, we measure the proportion of male and female actively exposed to deviant content (by following deviant account) broken down by age (Figure 16).

Surprisingly both in Tumblr and in Flickr, despite the differences in the demographic of the users and the typology of social platform, we observe a similar trend. The curve for men shows an increasing trend that plateaus at its maximum in the range of age 40 to 55. In contrast, women, although less exposed than men at any age, have their peak in their 20s, much earlier than men. This peak is longer in Flickr up to 25 probably because the platform target a more adult audiance. This observation supports previous findings [30] and explains the distributions we observed. In absolute terms the volume of users consuming adult content (by following) in relation to the gender with different ages is interesting to compute and this is one of the first large-scale study shedding light on that. In Figure 17 we report the ratio male users over female users for different age ranges for adult content consumers both in Tumblr and in Flickr. If we compare those values with the general ratio of the population of the whole online social network we discover a similar trend. The consumption of adult content is substantially equal between the two genders under 25 years old. After that age the percentage of male users increases progressively compared to the female users. In particular in Flickr which is an platform targeting more adult people, the male consumption keep increasing until the age of 50.

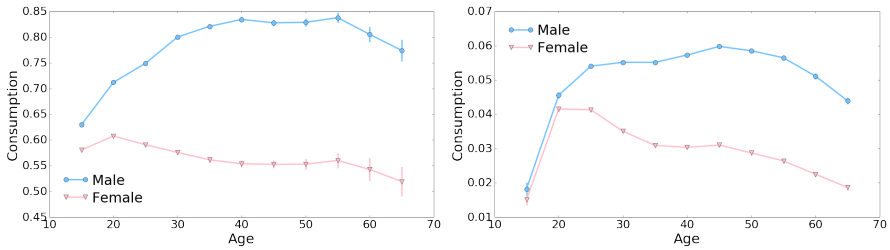

Fig. 16: Ratio of male and female consuming adult content for different age bands in Tumblr (left) and in Flickr (right).

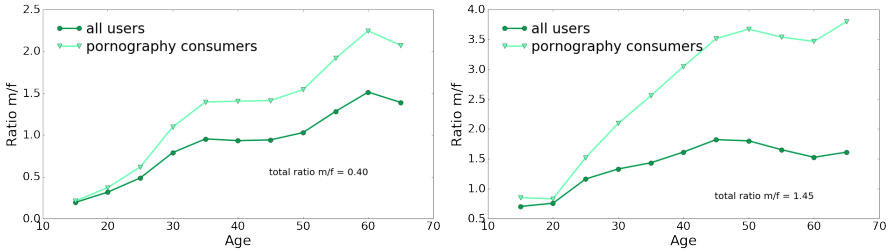

Fig. 17: Ratio of male and female users consuming adult content for different age bands compared to the ratio of the general population of the OSN in the sample for Tumblr (left) and Flickr (right).

Q9) What is the cross-platform behavior of deviant content consumers?

To better understand the dynamics of diffusion of deviant content, one could consider how adult material leaks from an online platform to another. It is very hard to track individual pieces of information moving across different social media layers. For example, a Flickr photo can be re-uploaded to Tumblr, which makes it difficult to trace it back to its original Flickr URL. To investigate how much the activity of people involved in the activity of production and consumption of adult content is cross-platform, we resort to an analysis at user-level instead. We do so by considering those users who have subscribed to both Flickr and Tumblr using the same email address. We found 293K users with an account in both platform in our sample. As shown in Figure 18, approximately 32% of them are involved in pornography at least in one of two networks, the majority using Flickr instead of Tumblr.

There are two competing hypothesis regarding the involvement of users on multiple social networks. On one hand, the literature in multiplex networks seems to point towards a scenario in which the activity indicators of nodes on multiple layers are storngly correlated [55]. This finding holds mostly for structural indicators (e.g., node degree), but if the same principle applied also to the *type* of activity, people who actively engage in activities connected to deviant content on one platform would likely do the same also on the other one. On the other hand, social science studies suggest that the role people plays is strongly dependent from the context they are acting in. Social identification depends only partially by the characteristics of the individual

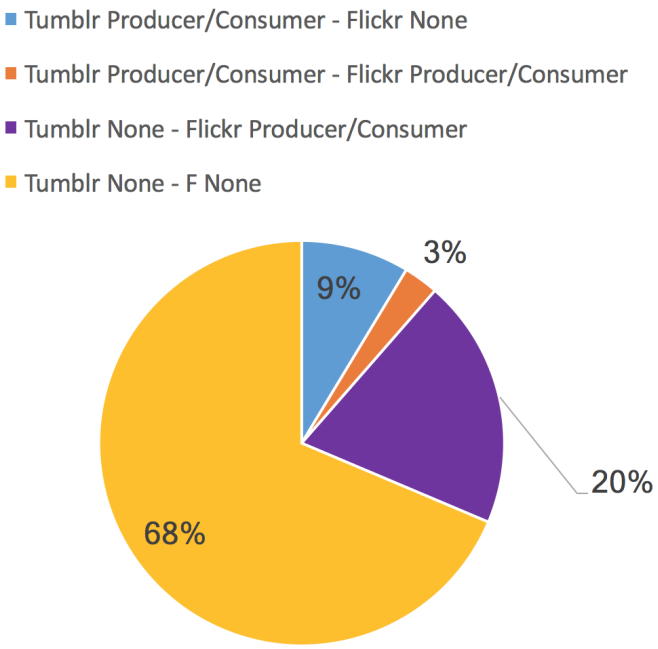

Fig. 18: Cross-platform behavior: matching users in Tumblr an in Flickr. Users are characterized by their role according to our classification (adult content producers or consumers. People not related to pornography even though they might be unintentionally exposed are classified as None.

and it is determined largely by factors that are distinctive of the group in which the individual is part [6, 68]. This is holds when comparing the behavior of an individual across groups and, consequently, across social systems. The fact that both Tumbr and Flickr do not enforce strong user identification (pseudonyms are allowed), makes it even easier for a person to build different, possibly contrasting online personas. This leads us to hypothesize that deviant users on one platform do not have the same type of behavior, on average, on the other.

In agreement with this second hypothesis, we find that only 3% (see Figure 18) of the users with matching account consume or produce deviant content in both platforms: people tend to use different social media for different purposes. Moreover, even when people use both platforms to deal with porno-graphical content their roles are often different (producing or consuming). We further investigated the users producing deviant content and we found that less than $1‰$ are producers in both networks.

5 Conclusion

This work aims at motivating researchers who study all types of deviant communities online as well as offline to explore in more depth the interaction between the agents in such networks and the external social environment. Our contribution scratches only the surface of the exploration space that underlies the many types of deviant networks and the multitude of settings they are situated within. The study we have presented is limited under many aspects, beginning from the focus on a single type of deviant behavior –adult material consumption– that is much more pervasive than others (e.g., anorexia) and, in that, has unique characteristics that likely cannot generalize to other deviant groups. In terms of methodology, alternative techniques (e.g., computer vision) could be used to identify adult content without a dedicated dictionary; those could possibly lead to describe the same phenomenon from a slightly different angle, for instance considering more exhaustively nodes that are not reached by search traffic or by tags. Furthermore the notion of pornographic content is culture-dependent; then it would be possible to study deviant behaviors under different cultural premises and not only under a western country perspective that we adopted.

Our work has limitations that leave space for exploration in future work. We have focused our study on Tumblr and Flickr, but more social media ecosystems could be included; Twitter would be a good cadidate as it does not enforce any strong restriction on the content of tweets. We have studied the spread of pornographic content, but a broader exploration of how other types of deviant content diffuses on general-purpose social media would be very desirable. For what concerns the analysis tools we have used to measure the importance that nodes have in spreading information, one could investigate a plethora of alternative ways to assess the role of groups and individuals in the process. In particular, we believe an interesting extension to our work would be to look at more sophisticated centrality metrics such as routing betweenness centrality [24].

Despite these limitations, we believe that our study has already important theoretical implications in revealing, for the first time on very large scale, that deviant communities can be deeply rooted into the relational fabric of a social network, and that the echo of their abnormal activity can reach a plenitude of ordinary users. Also, from a practical point of view, learning the effect that a minority group can have on a much larger audience is key to trigger mechanisms able to contain risky deviant phenomena by means of targeted interventions on few nodes, as we have shown. We believe that this work could set the basis for a line of study that could lead to a deeper understanding of deviant networks and of their impact on everyone's life.

References

1. Adamic, L. A. and Glance, N. (2005). The political blogosphere and the 2004 us election: divided they blog. In *International workshop on Link discovery*. ACM.
2. Aiello, L. M. (2015). Group types in social media. In Paliouras, G., Papadopoulos, S., Vogiatzis, D., and Kompatsiaris, Y., editors, *User Community Discovery*,

Human–Computer Interaction Series, pages 97–134. Springer International Publishing.
3. Aiello, L. M., Barrat, A., Cattuto, C., Ruffo, G., and Schifanella, R. (2010). Link creation and profile alignment in the aNobii social network. In *SocialCom*.
4. Aiello, L. M., Barrat, A., Schifanella, R., Cattuto, C., Markines, B., and Menczer, F. (2012). Friendship prediction and homophily in social media. *ACM Transactions on the Web*, 6(2):9:1–9:33.
5. Allen, M., D'Alessio, D., and Brezgel, K. (1995). A meta-analysis summarizing the effects of pornography ii aggression after exposure. *Human communication research*, 22(2):258–283.
6. Ashforth, B. E. and Mael, F. (1989). Social identity theory and the organization. *Academy of management review*, 14(1):20–39.
7. Attwood, F. (2005). What do people do with porn? qualitative research into the comsumption, use, and experience of pornography and other sexually explicit media. *Sexuality and culture*, 9(2).
8. Barbieri, N., Bonchi, F., and Manco, G. (2013). Cascade-based community detection. In *WSDM*. ACM.
9. Bessi, A., Coletto, M., Davidescu, G. A., Scala, A., Caldarelli, G., and Quattrociocchi, W. (2015). Science vs conspiracy: collective narratives in the age of misinformation. *PloS one*, 10(2):02.
10. Blackburn, J., Simha, R., Kourtellis, N., Zuo, X., Ripeanu, M., Skvoretz, J., and Iamnitchi, A. (2012). Branded with a scarlet c: Cheaters in a gaming social network. In *WWW: Proceedings of the 21st International Conference on World Wide Web*, pages 81–90, New York, NY, USA. ACM.
11. 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 Experiment*, 2008(10):P10008.
12. Boero, N. and Pascoe, C. J. (2012). Pro-anorexia communities and online interaction: Bringing the pro-ana body online. *Body & Society*, 18(2):27–57.
13. Buzzell, T. (2005). Demographic characteristics of persons using pornography in three technological contexts. *Sexuality & Culture*, 9(1):28–48.
14. Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, P. K. (2010). Measuring user influence in twitter: The million follower fallacy. *Icwsm*, 10(10-17):30.
15. Chen, A.-S., Leung, M., Chen, C.-H., and Yang, S. C. (2013). Exposure to internet pornography among taiwanese adolescents. *Social Behavior and Personality: an international journal*, 41(1):157–164.
16. Cheng, J., Adamic, L., Dow, P. A., Kleinberg, J. M., and Leskovec, J. (2014). Can cascades be predicted? In *Proceedings of the 23rd international conference on World wide web*, pages 925–936. ACM.
17. Christakis, N. A. and Fowler, J. H. (2008). The collective dynamics of smoking in a large social network. *New England journal of medicine*, 358(21):2249–2258.
18. Clinard, M. and Meier, R. (2015). *Sociology of deviant behavior*. Wadsworth Cengage Learning.
19. Coletto, M., Aiello, L. M., Lucchese, C., and Silvestri, F. (2016). On the behaviour of deviant communities in online social networks. In *ICWSM 2016, Köln, Germany*.

20. Conover, M., Ratkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F., and Flammini, A. (2011). Political polarization on twitter. In *ICWSM*.
21. Davis, J. P. (2002). The experience of 'bad'behavior in online social spaces: A survey of online users. *Social Computing Group, Microsoft Research*.
22. De Choudhury, M. (2015). Anorexia on tumblr: A characterization study. In *Digital Health*. ACM.
23. Dolev, S., Elovici, Y., and Puzis, R. (2010a). Routing betweenness centrality. *Journal of the ACM (JACM)*, 57(4):25.
24. Dolev, S., Elovici, Y., and Puzis, R. (2010b). Routing betweenness centrality. *Journal of ACM*, 57(4).
25. Dow, P. A., Adamic, L. A., and Friggeri, A. (2013). The anatomy of large facebook cascades. In *ICWSM*.
26. Dunlop, P. D. and Lee, K. (2004). Workplace deviance, organizational citizenship behavior, and business unit performance: The bad apples do spoil the whole barrel. *Journal of organizational behavior*, 25(1):67–80.
27. Easley, D. and Kleinberg, J. (2010). *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press.
28. Feld, S. L. (1991). Why your friends have more friends than you do. *American Journal of Sociology*, pages 1464–1477.
29. Feller, A., Kuhnert, M., Sprenger, T. O., and Welpe, I. M. (2011). Divided they tweet: The network structure of political microbloggers and discussion topics. In *ICWSM*.
30. Ferree, M. (2003). Women and the web: Cybersex activity and implications. *Sexual and Relationship Therapy*, 18(3):385–393.
31. Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3):75–174.
32. Gavin, J., Rodham, K., and Poyer, H. (2008). The presentation of "pro-anorexia" in online group interactions. *Qualitative Health Research*, 18(3):325–333.
33. Grabowicz, P. A., Aiello, L. M., Eguiluz, V. M., and Jaimes, A. (2013). Distinguishing topical and social groups based on common identity and bond theory. In *WSDM*. ACM.
34. Guerra, P. H. C., Meira Jr, W., Cardie, C., and Kleinberg, R. (2013). A measure of polarization on social media networks based on community boundaries. In *ICWSM*.
35. Haas, S. M., Irr, M. E., Jennings, N. A., and Wagner, L. M. (2010). Online negative enabling support groups. *New Media & Society*.
36. Hald, G. M. (2006). Gender differences in pornography consumption among young heterosexual danish adults. *Archives of sexual behavior*, 35(5):577–585.
37. Hald, G. M., Malamuth, N. N., and Lange, T. (2013). Pornography and sexist attitudes among heterosexuals. *Journal of Communication*, 63(4):638–660.
38. Hald, G. M. and Štulhofer, A. (2015). What types of pornography do people use and do they cluster? assessing types and categories of pornography consumption in a large-scale online sample. *The Journal of Sex Research*, pages 1–11.
39. Hodas, N. O., Kooti, F., and Lerman, K. (2013). Friendship paradox redux: Your friends are more interesting than you. In *ICWSM*.
40. Kayes, I., Kourtellis, N., Quercia, D., Iamnitchi, A., and Bonchi, F. (2015). The social world of content abusers in community question answering. In *WWW: Pro-*

ceedings of the 24th International Conference on World Wide Web, pages 570–580, New York, NY, USA. ACM.
41. Kitsak, M., Gallos, L. K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H. E., and Makse, H. A. (2010). Identification of influential spreaders in complex networks. *Nature physics*, 6(11):888–893.
42. Kühn, S. and Gallinat, J. (2014). Brain structure and functional connectivity associated with pornography consumption: the brain on porn. *JAMA psychiatry*, 71(7):827–834.
43. Kvalem, I. L., Træen, B., Lewin, B., and Štulhofer, A. (2014). Self-perceived effects of internet pornography use, genital appearance satisfaction, and sexual self-esteem among young scandinavian adults. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 8(4).
44. LaSala, R. (2012). The social makeup of social media. *Compete.com Tech Blog*.
45. Lee, L.-H. and Chen, H.-H. (2011). Collaborative cyberporn filtering with collective intelligence. In *SIGIR*. ACM.
46. Lerman, K., Yan, X., and Wu, X.-Z. (2016). The "majority illusion" in social networks. *PLoS ONE*, 11(2):1–13.
47. Leskovec, J. and Horvitz, E. (2008). Planetary-scale views on a large instant-messaging network. In *Proceedings of the 17th international conference on World Wide Web*, pages 915–924. ACM.
48. Leskovec, J., Kleinberg, J., and Faloutsos, C. (2005). Graphs over time: densification laws, shrinking diameters and possible explanations. In *ACM SIGKDD*. ACM.
49. Leskovec, J., McGlohon, M., Faloutsos, C., Glance, N. S., and Hurst, M. (2007). Patterns of cascading behavior in large blog graphs. In *SDM*, volume 7, pages 551–556. SIAM.
50. Martin-Borregon, D., Aiello, L. M., Grabowicz, P., Jaimes, A., and Baeza-Yates, R. (2014). Characterization of online groups along space, time, and social dimensions. *EPJ Data Science*, 3(1):8.
51. Mitchell, K. J., Finkelhor, D., and Wolak, J. (2003). The exposure of youth to unwanted sexual material on the internet a national survey of risk, impact, and prevention. *Youth & Society*, 34(3):330–358.
52. Morgan, E. M., Snelson, C., and Elison-Bowers, P. (2010). Image and video disclosure of substance use on social media websites. *Computers in Human Behavior*, 26(6):1405–1411.
53. Negoescu, R. A. and Gatica-Perez, D. (2008). Analyzing flickr groups. In *CIVR*, New York, NY, USA. ACM.
54. Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582.
55. Nicosia, V. and Latora, V. (2015). Measuring and modeling correlations in multiplex networks. *Physical Review E*, 92(3):032805.
56. Pastor-Satorras, R. and Vespignani, A. (2005). Epidemics and immunization in scale-free networks. *Handbook of graphs and networks: from the genome to the internet*, pages 111–130.
57. Phillips, D. J. (1996). Defending the boundaries: Identifying and countering threats in a usenet newsgroup. *The information society*, 12(1):39–62.

58. Pingdom (2012). Report: Social network demographics in 2012. *Pingdom.com Tech Blog*.
59. Ramos, J. d. S., Pereira Neto, A. d. F., and Bagrichevsky, M. (2011). Pro-anorexia cultural identity: characteristics of a lifestyle in a virtual community. *Interface (Botucatu)*, 15(37):447–460.
60. Ratkiewicz, J., Conover, M., Meiss, M., Gonçalves, B., Patil, S., Flammini, A., and Menczer, F. (2011). Detecting and tracking the spread of astroturf memes in microblog streams. In *WWW*.
61. Romero, D. M., Tan, C., and Ugander, J. (2013). On the interplay between social and topical structure. In *ICWSM*.
62. Romito, P. and Beltramini, L. (2015). Factors associated with exposure to violent or degrading pornography among high school students. *The Journal of School Nursing*, 31(4):280–90.
63. Sabina, C., Wolak, J., and Finkelhor, D. (2008). The nature and dynamics of internet pornography exposure for youth. *CyberPshychology & Behavior*, 11(6).
64. Schifanella, R., Barrat, A., Cattuto, C., Markines, B., and Menczer, F. (2010). Folks in folksonomies: Social link prediction from shared metadata. In *WSDM*. ACM.
65. Schuhmacher, M., Zirn, C., and Völker, J. (2013). Exploring youporn categories, tags, and nicknames for pleasant recommendations. In *Workshop on Search and Exploration of X-Rated Information*. ACM.
66. Shores, K. B., He, Y., Swanenburg, K. L., Kraut, R., and Riedl, J. (2014). The identification of deviance and its impact on retention in a multiplayer game. In *CSCW: Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 1356–1365, New York, NY, USA. ACM.
67. Suler, J. R. and Phillips, W. L. (1998). The bad boys of cyberspace: Deviant behavior in a multimedia chat community. *CyberPsychology & Behavior*, 1(3):275–294.
68. Thibaut, J. W. and Kelley, H. H. (1959). The social psychology of groups. *Database: PsycINFO*.
69. Tyson, G., Elkhatib, Y., Sastry, N., and Uhlig, S. (2013). Demystifying porn 2.0: A look into a major adult video streaming website. In *IMC*. ACM.
70. Tyson, G., Elkhatib, Y., Sastry, N., and Uhlig, S. (2015). Are People Really Social in Porn 2.0? In *ICWSM*.
71. Wasserman, S. and Faust, K. (1994). *Social network analysis: Methods and applications*, volume 8. Cambridge university press.
72. Wellen, J. M. and Neale, M. (2006). Deviance, self-typicality, and group cohesion the corrosive effects of the bad apples on the barrel. *Small Group Research*, 37(2):165–186.
73. Wolak, J., Mitchell, K., and Finkelhor, D. (2007). Unwanted and wanted exposure to online pornography in a national sample of youth internet users. *Pediatrics*, 119(2):247–257.
74. Xu, J. and Chen, H. (2008). The topology of dark networks. *Communications of the ACM*, 51(10):58–65.
75. Ybarra, M. L. and Mitchell, K. J. (2005). Exposure to internet pornography among children and adolescents: A national survey. *CyberPsychology & Behavior*,

8(5):473–486.