Breaking bad online: a synthesis of the darker sides of social networking sites

Dionysios Demetis

This essay deconstructs the ultra-dark side of social media and explores the variety of 'bad' behaviour online by looking at a wide spectrum of exploitative practices. Through the use of primary data from an online platform, we posit the question 'What's the worst thing you've done online'? We collect, code and synthesise the fully anonymised discussions and develop a classification model for bad online behaviour. We combine the categories that emerge from our empirical data with those proposed by Baccarella et al. (2018) and develop a new combined (meta-) classification model that captures both the dark side of social networking and the ultra-dark. A framework is proposed for conceptualising the spectrum of exploitative practices and the essay concludes by providing a series of management considerations.

Keywords: social media, social networking, dark side, cybersecurity, cybercrime

INTRODUCTION

With the development of social networking sites (SNSs), an information ecosystem has emerged that has become highly complex. This rich ecosystem allows us to share information and gives rise to a variety of social media activities and phenomena (Kietzmann et al., 2011). There is no doubt that the interconnectedness enabled by SNSs is both a foundational aspect of their success and a springboard for the richness of activities that they support. In fact, the value of such networks has been expressed early on by Robert Metcalfe in the context of telecommunications. Metcalfe estimated that the value of a network is proportional to the square of the number of its nodes, as more nodes allow for a non-linear progression of interactions. Put simply, the added value lies in the possibilities to interconnect and in the 'space between the nodes', not in the number of nodes *per se*; more recently, empirical data from Tencent and Facebook have validated Metcalfe's Law (Zhang et al., 2015).

But, while the value of SNSs has been increasing through an expanding network of nodes and interactions, a number of negative phenomena has emerged. These can be conceived collectively as the *dark side* of social media (Baccarella et al., 2018). Mistrust, deception, exploitation, fraud and privacy violations are few of its many expressions. These place individuals directly at risk (e.g. blackmail), and cast societies into new territories of concern (e.g. election influence through SNSs). Phenomena that express the dark side, not only jeopardise the potential of SNSs, but can also escalate to more grave phenomena at the macro-level. Such phenomena diffuse at an accelerated rate. For example, in research published in the journal *Science*, it was found that false news spreads 'significantly farther, faster, deeper and more broadly' than true news (Vosoughi et al., 2018, p. 1). Ironically, misinformation is perceived to be more interesting (because of its novelty) than factual information. Ultimately, the scope and spread of the dark side of SNSs will demand much closer attention from scholars and social media practitioners; this paper is motivated by that need and concentrates on developing a classification model that encapsulates darker shades of SNSs.

RELATED WORK

The explosion of the dark side of SNSs can be considered at two main levels (the micro level and the macro level). At the micro level, negative side effects abound. From trolling (Synnott et al., 2017) to online radicalisation (Omotoyinbo, 2014) to cyberstalking (Spitzberg & Hoobler, 2002), SNSs enable the reconstruction of known negative social effects into the online space. The micro level then acts as a springboard for the development of macro-level effects. While individual users operate within the confines of their own interconnected social spheres, they are always prompted to step outside of that sphere and interconnect more. This leads to network effects and to a progressive development of more interconnected networks (Capra & Luisi, 2014).

In Facebook, for example, this has led to a rather steady and progressive expansion to 2.45 billion monthly active users as of the third quarter of 2019 (Clement, 2019). The increasing interconnectedness experienced in SNSs leads to an entropic trend for more significant macro-scale phenomena (Letichevsky et al., 2017). With SNSs having reached a deeper social integration and becoming more and more interwoven with the spectrum of socio-economic and political affairs, new phenomena surface (e.g. election influence, large-scale privacy and data breaches).

Of course, the early online development of these social side effects is hardly surprising within the dark side of SNSs; cybercriminals are almost always early adopters of new technology, along with tactics that will allow them to avoid detection (McMurdie, 2018). Thus, the vandalising of the information society is a natural extension of social/antisocial or legal/illegal realities, with cybercrime and other negative consequences of SNSs being an ‘inevitable downside of the information society’ (Furnell, 2003, p. 8).

Consequently, the study of the dark side of SNSs is critical but research to that end is still at an embryonic stage, concentrating mostly on trolling and other behavioural phenomena, which constitute online variants of harassment (Craker & March, 2016). One important step towards exploring the structure of the dark side, and the way in which it is expressed through different phenomena is the adapted honeycomb framework for the dark side of social media (Baccarella et al., 2018). This is based on an earlier classification approach where social media functionality rests on a few key pillars: conversations, sharing, presence, identity, relationships, groups and reputation (Kietzmann et al., 2011). Through their framework, Baccarella et al. (2018) propose that each identified pillar of social media can give rise to a dark side expression. For conversations, we have misinformation, disinformation and aggressive engagement. For identity, the exploitation of online self while sharing inappropriate content and distribution. For groups, it is in-group and out-group bias. For relationships, it is threat, coercion, abuse and intimation. For reputation, it is shaming and defamation and for presence, it is location tracking and monitoring. This is an interesting deconstruction that calls for empirical exploration and further consideration.

In their paper, Baccarella et al. (2018) prompt us to engage more with researching the dark side of social media, and to develop alternative ways of deconstructing these phenomena. Empirical verification of these dark side categories remains also of interest. Thus, the authors recognise that researchers need to explore the dark side further, as this is not adequately captured in existing research on the topic. Of course, research on the dark side does exist across different orientations and disciplines; for instance, scholars explore employee pressure in work-related social media use, where social media use is seen as a source of ‘boundary conflicts, causing spillover effects across life domains, which in turn are associated with exhaustion’ (van Zoonen et al., 2016, p. 19). Similarly, while some significant work has been done on some extremist- and terrorism-related topics on social media (Johnson, 2019), the darkest side of social media is much broader and remains largely unexplored (Salo et al., 2018). Between terrorism and online harassment, there remains a large spectrum of dark and ultra-dark activities that must be explored. Thus far, the spectrum of activities on the darker sides of the social networking spectrum has mostly been the subject of attention by psychologists, who attempt to delineate the behavioural and personality traits of those that engage with the dark side and might exhibit addictions, sadistic impulses and other similar phenomena (Kircaburun & Griffiths, 2018). Still, given the significance of the topic and a lack of exploring the ultra-dark side of social networks, this essay concentrates on delineating the characteristics of the spectrum of activities in the darkest side of SNSs.

In this context, the research mentioned above illustrates the need to do more work to delineate the structure of the darker side of SNSs before the deployment of countermeasures can be considered. Thus, before any prevention and countermeasure strategies can be considered, we argue that there is a pressing need to dive first into the deconstruction of the darker sides of SNSs, and realise that all such phenomena have a spectrum (Whittle et al., 2013). As there are many shades of the dark side of SNSs, this paper tries to combine the categorical assumptions behind phenomena of the dark side as proposed by Baccarella et al. (2018), with phenomena of the ultra-dark side; the goal is to develop a model that brings both together.

At the core of the dynamics of the darker sides of SNSs are negative effects, and these are usually discussed in the context of the more widespread SNSs (e.g. Facebook, Instagram and Twitter); however, we must remember that all cyberspaces can be avenues for exploitation (Dhillon & Backhouse, 2001). Naturally, professional SNSs are not exempt. In a study conducted by Silic and Back (2016) on LinkedIn, the findings indicate that users face significant and hardwired obstacles in avoiding exploitative behaviour. Individuals are easily subjected to deception and victimisation because attackers manage to tap into some of the innate characteristics that all humans share. By exploiting such characteristics, attackers manage to elicit a psychological reaction from the users instead of a rational one. These reactions relate to how we respond as humans when we sense (online) danger. In such circumstances, the part of our brain called the *amygdala* takes over (Davis & Whalen, 2001). The 'amygdala's job is to quickly process and express emotions, especially anger and fear. This little mass of grey matter is the watchdog of the brain, always remaining alert for times we might be threatened. When it does sense danger, it can completely take over, or hijack the upstairs brain. That's what allows us to *act* before we *think'* (Bluth & Blanton, 2015, p. 42). Unfortunately, for an online user that is undergoing trolling, online extortion or any other phenomenon of the (ultra-) dark side, this contingency becomes problematic. It is the mechanism that can prompt users to online actions that are compromising. The same applies to organisations that are also users of SNSs (Roshan et al., 2016).

While there is a neural signature on social norm compliance that is an important basis for the development of human sociality itself (Spitzer et al., 2007), it is equally important to recognise that in the darker sides of SNSs, the '*act* before *think'* innate behavioural characteristic that humans share is often what gets users into trouble. In an interview with former cybercriminal Brett Johnson who was labelled by the United States Secret Service as the 'Original Internet Godfather' and spent six years in prison that barred him from computer use, it becomes clear that online deception can deviate from calculated attempt towards user exploitation. As Johnson describes from his years as a cybercriminal (now a cybersecurity consultant and keynote speaker), deception is also an art because social engineers will pick up really subtle signals that they will then use. Not only do they know how to collect data and exploit their victims by pushing just the right behavioural buttons, but they also collaborate feverishly amongst themselves and share techniques, insights and exploitation tactics (Demetis, 2018a). Once a victim is exploited online, the technique of exploitation is shared online amongst those with a similar interest, and this gives rise to a distributed cyber-criminality where: a) many participants will target the same victim and b) the technique of exploitation will be reused variably to target new victims. Thus, the same technique of exploitation will reappear in different contexts, variations and applications, raising the variety of the dark side and the spectrum of its applications. In that context, SNSs have become 'important security holes where, with the use of social engineering techniques, malicious attacks are easily facilitated' (Silic & Back, 2016, p. 35).

Of course, there are several other side effects. For example, the study of negative psychological and relational experiences on Facebook through focus groups reveals further interesting trends. In their work, Fox and Moreland (2015) uncover five themes as SNSs stressors: managing inappropriate or annoying content, being tethered, lack of privacy and control, social comparison and jealousy and relationship tension and conflict. Yet, despite the spectrum of negative emotions experienced in SNSs, users return to these platforms because of their fear of missing out, keeping up with content and peer pressure. Coupled with the deep penetration of mobile computing, the 'excessive usage and habitual checking (of phones) may result in compulsive usage and even lead to mobile phone addiction…the smartphone user's increased experience of technostress will cause greater feelings of stress for that user', so negative and unintended consequences can intensify (Lee et al., 2014). In many cases, SNSs have become online platforms where the dark tetrad personality traits (i.e. narcissism, Machiavellianism, psychopathy and sadism) can be allowed to flourish (Craker & March, 2016). In fact, the dark side of online environments, has seen such an explosion of negative behaviour that it has led to the emergence of the novel field of *cyberpsychology* (Barton, 2016). With such a wide spectrum of activities that covers the dark side of SNSs, one central question is how we can begin to deconstruct these phenomena in a way that elevates the degree of their generalisability (Lee & Baskerville, 2003).

For this reason, we need to develop a synthesis of the dark and the ultra-dark sides of SNSs and assimilate new phenomena that can expand the categorical assumptions being made. Towards this goal, the categories that are described by Baccarella et al. (2018) form a basis upon which we can look into the ultra-dark side of the spectrum of SNSs.

The synthesis attempted in this essay is informed by empirical data from 84 observations from an online platform (Omegle), where users are called upon to discuss what is the 'worst thing that they have done online' under complete anonymity that is offered by the platform (this does not mean that these activities took place on the platform itself that displays plenty of precautions). The remainder of this essay is organised as follows: the next section introduces the methodology and then the discussion section presents the key findings, a mind-map of the interactions/activities delineated, and then concentrates on their synthesis with the categories identified by Baccarella et al. (2018) before presenting an overarching model.

METHODOLOGY

The research follows an interpretivist epistemology (Klein & Myers, 1999; Walsham, 1995) and an inductive reasoning. The goal of the research was to explore the *variety* of 'bad' behaviour online and the darker spectrum of options for bad behaviour that users can engage with. In contexts that have not been explored as much, like in the case of the dark side of social media, the decision was taken to map the variety of activities first, before connecting them to the categories developed by Baccarella et al. (2018). This approach allows us to widen the net of exploration and combine the dark side of SNSs with the ultra-dark side, ultimately leading to the development of an encompassing classification framework.

To collect primary data, the Omegle platform was accessed over a period of two months. The platform itself is branded as giving the opportunity to ‘talk to strangers’ and secures anonymity. The platform displayed plenty of clear and visible notices upfront to users, prompting them to be careful online and that the platform could be exploited/misused for predatory purposes. Users were also notified that they should stay on high alert. As indicated by the platform itself:

> ‘When you use Omegle, we pick someone else at random and let you talk one-on-one. To help you stay safe, chats are anonymous unless you tell someone who you are (not suggested!), and you can stop a chat at any time. Predators have been known to use Omegle, so please be careful’.

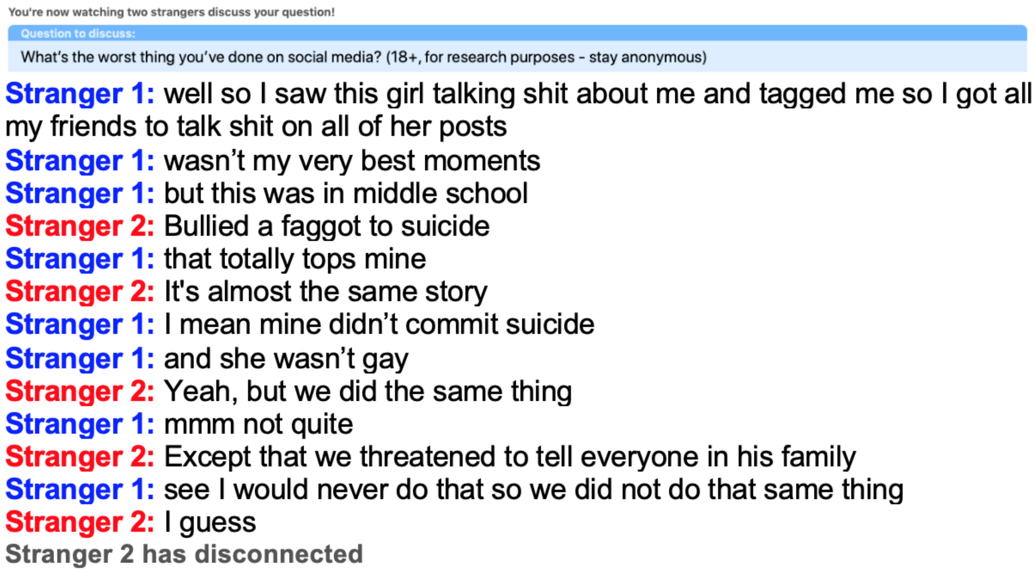
More importantly for the purposes of this research, the platform allows different users to interconnect and discuss a question that is posed to them by a third party (the researcher in this case). The users are both aware that they are discussing a question posed by someone else, and that they are being observed during their discussion. They see each other as ‘Stranger 1’ and ‘Stranger 2’, and all identities are hidden while they are paired into the discussion randomly. Before being paired online, the users have to click on a button below a paragraph that says they ‘volunteer to be watched’ and the observer/researcher gets the notification that ‘You’re now watching two strangers discuss your question’. For the purposes of this paper, the main question that was posed to the users was: ‘*What’s the worst thing you’ve done on social media’?* and variants of questions that would discuss bad behaviour perceived like ‘*What bad behaviour have you encountered on social media’? or ‘What’s the worst thing you’ve seen or done online’?*

Questions posed were accompanied with an indication that this is for an 18+ audience and for research purposes. Unfortunately, in about 98% of the cases, discussants would exit the conversation almost immediately and within a few seconds. Naturally, even if one of the discussants exited the online room, the text-based discussion would be terminated completely for all remaining parties and participants would be disconnected from the online room. This made the process of data collection very time-consuming, but it is understandable that many opted out from the discussion. Thus, repetitive initiations of online chats were required through the platform as the overwhelming majority of users would disconnect, either because they would choose to opt out of discussing the question itself, or because at the very beginning of the conversation with the stranger they were paired with, there was no real interaction or communication at a level that would be agreeable to both users. Another complicating factor is that bots would often join the chats, so users would log out immediately after realising that.

It is important to note here a few additional considerations: while the framing of the question to the users involved *all social media* and it did not specify any particular platform, discussants spoke generally of such behaviour and occasionally mentioned Facebook, Instagram, KikMe and a few more SNSs. The findings reflected in this essay are not platform-specific. The variety of the phenomena discussed and the approaches taken are more important than the specific occurrences. In this regard, the activities that were discussed by the participants (bullying, exploitation, fraud, extortion, etc.) were (mis) uses of all SNSs and not of the Omegle platform *per se.* In fact, from a cybersecurity perspective it could be argued that Omegle is safer than other SNSs as it is: *anonymous by default* at the point of entry; it does not allow for the *exchange of files* (which could contain security threats in themselves like malware), has appropriate notifications on its homepage and *no login/account functionality* for establishing stable cross-party interactions that would maximise exposure over time. Naturally, no claims are being made towards the exploitation of the platform itself and none were observed. Indeed, the platform seems to have taken ample precautions, and there are visible warnings for the users.

Video discussions were not initiated and the text-based discussion was selected instead; these were logged with a serial number of the discussion. This is depicted as #n where n is the serial number given to that discussion (e.g. #45 would be referenced in the discussion, if a quotation from a user would be used from that corresponding text log). With the fear of stating the obvious, the veracity of individual users' responses can be called into question, though that is a common trait in all computer-mediated communication that affords anonymity; under such circumstances of platform-secured anonymity, it is recognised that you can have simultaneous positive and negative effects depending on who is participating in online interactions and what groups form relations between them (e.g. white supremacists that can use SNSs for propagating racist material versus minorities like homosexuals that may network in SNSs with the ultimate goal of changing the laws in the country) (Christopherson, 2007).

Of course, as this essay concentrates on the darker side of the spectrum, it is difficult to take the utterances by strangers that are secured by anonymity as a true testament of what had actually occurred. This is a limitation that is recognised in all online research that rests on anonymity (Lefever et al., 2007). However, the very *utterance of an act* also raises a conditioning for the *possibility of its actualisation*: either by the same or other users. Thus, while the association between user and act cannot be verified in conditions of anonymity (unless one conducts interviews with convicted criminals that would represent a much narrower scope of activity), one can collect a variety of techniques/actions that are likely to occur.



You're now watching two strangers discuss your question!

Question to discuss:

What's the worst thing you've done on social media? (18+, for research purposes - stay anonymous)

Stranger 1: well so I saw this girl talking shit about me and tagged me so I got all my friends to talk shit on all of her posts
Stranger 1: wasn't my very best moments
Stranger 1: but this was in middle school
Stranger 2: Bullied a faggot to suicide
Stranger 1: that totally tops mine
Stranger 2: It's almost the same story
Stranger 1: I mean mine didn't commit suicide
Stranger 1: and she wasn't gay
Stranger 2: Yeah, but we did the same thing
Stranger 1: mmm not quite
Stranger 2: Except that we threatened to tell everyone in his family
Stranger 1: see I would never do that so we did not do that same thing
Stranger 2: I guess
Stranger 2 has disconnected

Figure 1. Sample of discussion record from an observation

Transcripts of chats were logged and appended in a text document. Then, iterative mind-mapping was chosen as an appropriate coding technique (Mazza, 2009). The software used for mind-mapping (Coggle) had a timeline feature that allowed the researcher to go back in time for the development of the mind-map and explore changes, additions and modifications, and it provides a basic skeleton of the variety seen throughout the observations. Furthermore, the mind-mapping technique allowed for a more malleable way to develop iterative expansions of the model through additional observations. This made the coding of the data more engaging and allowed for the advancement of the description of the model. An alternative was considered (nVivo), but as pure text-based data coding methods offer a standard approach, *visual data coding* offers more malleability, particularly in cases where the visual depiction of text requires flexibility for the emergence of interpretations (Mazza, 2009; Spence, 2007). In simpler terms, visual coding was interactive and provided flexibility to reorganise categories easily or revert to former versions of coding through the timeline feature that kept back-ups of all former incremental versions. The evolving field of visualisation is exciting for scholars as technology now ‘allows users to explore the underlying data’ (Isenberg et al., 2011, p. 310). Benefits to visual data coding include: a) editing convenience when compared to traditional linear text coding analyses and the ability to reconfigure visual components (Křemen et al., 2012), b) possible theoretical expansion to a large number of branches/levels that, combined with editing, allows for several iterations during an inductive reasoning process, c) better interaction with the data (Spangler et al., 2002) and d) better focused learning leading to creating more effective cognitive constructs (Shams & Seitz, 2008).

Once the data collected were visualised, then another branch was created with the seven (7) elements identified by Baccarella et al. (2018) to represent their honeycomb framework of the dark side on *identity* (exploitation of online self), *sharing* (inappropriate content and distribution), *relationships* (threat, coercion, abuse and intimidation), *reputation* (shaming and defamation), *groups* (in-group and out-group bias), *presence* (location tracking and monitoring) and *conversations* (misinformation, disinformation and aggressive engagement). Information visualisation and visual coding were most critical at this stage, as the ability to move visual categories around for their interconnections to be progressively linked together was very important. Naturally, some overlap between the elements identified by Baccarella et al. (2018) and this essay was to be expected; at the same time, the exploration of the ultra-dark side of SNS has yielded several new elements for consideration. The section that follows discusses the findings before the interconnections with the Baccarella et al. (2018) framework are developed. The section after that (the discussion) provides a synthesis of the categories and considers managerial implications.

EMERGING CATEGORIES FOR THE ULTRA-DARK SIDE OF SNSs

The sheer variety of activities that users discussed among themselves was staggering. The dark side seems to be flourishing with a number of activities that create plenty of opportunities for users of SNSs and cover a wide range of interests. Before we proceed to unravel the more generalisable associations, it would be useful to see some of the key categories from the activities as they have been classified, as well as discuss some of their implications. The key categories that we find to be dominating the dark side of SNSs based on the empirical data collected are presented below in Table 1 and discussed immediately after.

(Ultra-)Dark side category	**Description and example**
Geolocation-related activities	The geographical location of a user can be compromised, exposed and used for various purposes (e.g. threats to physical harm can be made more targeted and location-based). Geolocation can open the possibility for uncovering the identity of an individual.
Online Child Sexual Exploitation (**online CSE**)	Possibly the darkest category discussed, this phenomenon involves one or more of the stages of exploitation in the participation, production and distribution of child pornographic imagery, mostly associated with the dark web.
Sex-related activities	Sex-related activities online have been present since the birth of the World Wide Web, but in the dark side of SNSs users may find themselves exploited or facing harm. These can also have a financial component (e.g. romance fraud).
Identity-related activities	Identity stands at the centre of attention as impersonation, catfishing, dark-hat fictional personas (e.g. pretending to be a murderer online) or deception-based white-hat personas (pretending to be a friend) all rely on some manipulation of identity. Also, identity masking and protection is important for those that have an offender role.
Financial activities	Attempting to defraud users by using a variety of different techniques.
Trading activities	Exchange-based related activities of digital or physical goods through online means (e.g. drug-related activities bought with cryptocurrencies, or trafficking of pornography and online CSE imagery) creates the opportunity for dark markets to emerge.
Dark-web enabled activities	Through elevated anonymity, early dark-web activities like drug trading in Silk Road are morphing into a series of phenomena
Direct threats and/or **time-sensitive** threats	Posing direct threats with a time-sensitive characteristic (e.g. post a bomb threat to a school, blackmail users and terrorist-related threats)
Subtle exploitation	Deploying deception-based exploitative tactics (e.g. convincing someone to take actions that could result in their own physical harm, and convincing someone to commit suicide)
Procurement-related activities	Hiring another party (e.g. a cybercriminal) to hack into an SNS account and conduct account takeover, where SNS-procured information is used for other forms of exploitation.
Gore-sharing activities	Sharing of extremely graphic videos and images on bestgore.com or similar networking sites where beheadings, executions, impalements and other acts of similar scope are being shared and commented upon.

Table 1. A summary of the categories uncovered

As one user framed it, SNSs (or at least its dark side) is a place where users can ‘suspend their humanity’ (#18). This is indeed a good description for the ultra-dark side of SNSs where admittedly, we find variable degrees of ‘humanity-suspension’ for different activities. In turn, the unintended consequences for those can vary, depending on how different users react to the social interactions they participate in. Of course, users can be lured to divulge sensitive and personal information about themselves through their own cyber actions, but also, become progressively desensitised from the online participation of such activities.

While every categorisation and classification remain problematic and an act of choice based on observer-relative characteristics that spawn the categories to begin with (Angell & Demetis, 2010), we accept that there are always different ways in which such categories can be drawn. This constitutes a limitation in any classification. More importantly, the categories identified above can be combined in launching more complicated attacks through SNSs. In exploring such combinations, users may perceive and interpret one activity (e.g. geolocation-related activity) as a characteristic of another activity (e.g. direct threat). These combinatory effects can increase the sophistication in the response required, or indeed, increase the level of cyber-awareness that users must maintain so that they steer clear of being exploited. For example, in one of the observations, the user said that they received a photograph with sexual content where the sender had not deleted the metadata. Then the recipient used geolocation-related exploitation by extracting the EXIF-metadata from an exchanged photograph and as she said, she ‘went on Google maps and sent him pictures of his house’ (#15) then his Facebook profile, other accounts and ‘blackmailed him for fun’ as the ‘chase was better than the catch, honestly’ (#15).

Of course, the combinatory possibilities that arise from the different elements and their perceived relations remain contextual. It also involves a dynamic of exploitation between one user and another (e.g. one user sends a photograph without thinking of metadata-based exploitation in the geographic-related activity, while the other user uses that vulnerability as an opportunity for further exploitation in the context of the online relationship).

We will discuss some implications for each category of Table 1 in brief:

1) **Geolocation**-related activities

In activities related to geolocation, we find that they become an input for other activities (e.g. blackmail, threats, lure to physical meetings under false identities where co-location is important, etc.). Naturally, where the possibility of physical harm arises then the dangers escalate and thus location-based characteristics are very important, and should also take priority when scholars are considering other classifications that would contain a risk-based dimension at their core. Several activities can be considered here in light of geo-social networks with photo sharing, friend tracking and 'check-ins' (Ruiz Vicente et al., 2011). In the observations, the extraction of metadata from photographs and location tracking (#15) was discussed, users being convinced to meetings under false identities (#24) and also IP tracking associated with threats (#9).

2) Online CSE (Child Sexual Exploitation)

One of the most serious activities in the context of the ultra-dark side of SNSs remains the online exploitation of children, which includes possession, distribution or the production of child pornography (Shelton et al., 2016). For this reason, this is included as a separate category here. Given the combinatory possibilities that we discussed previously, it is also enabled by the dark web as well. In the few occurrences that were observed in the discussions, trafficking of child images, requesting child pornography and predatory attempts to lure underage children were discussed, with users promptly leaving the discussion after saying 'I've sent children nudes' (#51) or 'watch child porn' (#23) or entering the discussion by saying 'hello.male here. looking for a girl around 15' (#20) and the other discussant responds by saying 'Hello paedophile' (#20). But with users being aware that the discussion is being monitored, they would tend to stop any online CSE-related discussion within seconds and logout; this is in contradistinction to the following category that attracted more discussion when it emerged as a topic between two participants.

3) Sex-related activities

Perhaps unsurprisingly, the importance of the internet as a medium for exploring human sexuality, and the emergence of cyber-sexuality as a space between fantasy and action (Ross, 2005) has given rise to various different forms through which users express themselves sexually through SNSs. In some cases, however, users may divulge information about themselves or others in a way that they are subjected to the dark side of SNSs, and many of these activities entail their own risks. Within that category, we can find the trading of procured pornography or of revenge porn, which seems to have become a growing plague (Stroud, 2014), leading to devastating privacy invasions. Despite the significance of such activities and their negative social impact, there seems to be only sporadic and disjunctive legislative responses (Goldsworthy et al., 2017). While the instances that were logged varied both in terms of 'humanity suspension' and potential risk to users, observations logged users discussing about: virtual cybersex-role playing (with users offering their usernames/account information on SNSs like KikMe), seeking sex with random strangers and social necro-porn; the risks involved include identity exposure/imagery, physical harm, etc., while the variety of cybersex-oriented risks remains much larger and entails a whole spectrum of security risks and legal liabilities (Chou et al., 2008).

4) Identity-related activities

The centrality of identity is another theme that emerged from the observations and has been shown to have a central significance in SNSs (Baccarella et al., 2018; Kietzmann et al., 2011). Also, the concept of an online identity has also been shown to maintain a multi-dimensional character. In a thorough study conducted by the *Future of Identity in the Information Society* (FIDIS) consortium, where the interactions between virtual persons have been mapped (FIDIS, 2009), it is shown that the variety of ontologies associated with virtual persons is surprising. In the context of this essay, the observations of users discussing themes related to identity involved: general impersonation, catfishing (intent to lure into a relationship), maintaining a 'black-hat' fictionate persona (pretending to be a murderer), or a 'white-hat' fictionate persona (pretending to be a friend) or even using any other covert identity role (e.g. cyberstalking). Sometimes identity-related activities have no discernible reason (other than the fact that they are possible in an online environment). For instance, in a discussion between two participants, Stranger 1 says: 'Pretended I was a murderer' and when Stranger 2 asks: 'why'? then Stranger 1 responds by saying: 'ldk' (I don't know) (#41). On other occasions, users were pretending to be racist (#55) or reversing a whole series of identity-related attributes (gender/age, etc.) so that they can support other interactions. As one user put it: 'I catfished someone when I was 11 saying I was

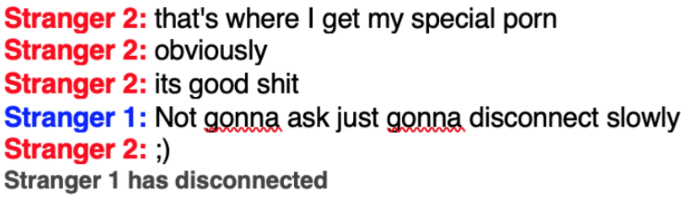
16' (#15). Of course, the identity reversals that we see online act as a preamble of attracting user attention, and play an important role across all categories.

5) Financial activities

For some users, financial motivations remain significant and SNSs allow users to explore a wealth of information that could be manipulated to conduct cyber-enabled fraud, though the scale of these activities and their related financial costs are a subject of debate and at an early stage of exploration (Anderson et al., 2013). Perhaps the most common characteristic that has been well known in the literature is the use of personal data from SNSs (e.g. date of birth, address, etc.) that are then used for financial fraud, applying for loans, credit cards, and so on (Demetis, 2018b; Lilley, 2000). A rarer encounter is micro-sums through SNSs interactions; for instance, one user narrated how they 'cheat someone and give me a paysafe for €20' (#39) though with the advent of cryptocurrencies, Bitcoin and other virtual assets have enabled a lot of the activity on the dark side.

6) Trading-related activities & **7) Dark-web** enabled activities

While trading-related activities may or may not include a financial/monetary exchange, they retain another special place in the dark side of SNSs. Whether it is physical or digital goods being traded, trading manages to relay a large number of activities in SNSs and reorients the nature of interactions. Or else, it creates a sustainable market that allows the dark side of SNSs to become darker (i.e. as an energy source through which the dark side feeds, preys upon and ultimately becomes darker). Because of that phenomenon, the dark side expands further, supported of course by the infinite replicability of digital data. Whether that is 'sending nudes' (#54), 'ordered illegal drugs' (#42) and 'went on Silk Road back in the day' (#34) or 'sent children nudes' (#51), the exchanges themselves create further pressure to participate in ongoing or developing social networks, with some users spelling out how that pressure is transferred to the users. In expecting a continuous flow of data for trading purposes, one user mentioned that if 'she keeps me happy with my demands, she won't have any accidents' (#15). Naturally, the sense of added security that users feel in the dark web due to the distribution and ping-ponging of IP addresses creates another comfort zone for some (see Figure 2).



Stranger 2: that's where I get my special porn
Stranger 2: obviously
Stranger 2: its good shit
Stranger 1: Not gonna ask just gonna disconnect slowly
Stranger 2: ;)
Stranger 1 has disconnected

Figure 2. Discussion on dark web theme with a possible online CSE link

8) **Direct threats and/or time-sensitive threats**

Sometimes, the time-horizon for exploitation is shrinking to the actualisation of direct threats, where the user (which could be an organisation) has little time to react. For example, one user claimed that he/she used SNSs, ‘made a bomb threat on my school’ and ‘got suspended for a mere week’ (#16). Others engage in blackmail and other activities that constitute direct threats to individuals and/or groups. While it was not observed during the discussions we logged, the exploitation of terrorist forums in promoting violence is also known (Walters, 2011) and can be classified in this category, as well as other threats of a similar nature.

9) **Subtle and indirect exploitation**

Of course, not all exploitation is direct. It can take more subtle forms of exploitation that progressively lead to another action. Through repetitive interactions that can be facilitated through SNSs and longer online-relationships, users can build more established personal relationships with others and these can lead onto more serious actions if they are ill-intended. One user for example mentions how he/she ‘probably made a girl kill herself’ (#47), another says that they ‘made fake accounts and reported b****** they hate’ (#71), while another takes a more active role and says that they ‘talked someone into committing suicide’ (#72). Online hate in its different expressions can be surprising; in fact, in another discussion, users (despite the question posed to them) would criticise such online behaviour, saying ‘I really can’t understand that hate…like’ ‘bring this b**** to commit suicide’ (#40).

10**) Procurement**-related activities and **account takeovers**

In a general cybercrime context, it is well known that users can procure the activities of cybercriminals by asking for specific services. However, the occurrences that were observed involved wild claims of hiring a hitman, or more straightforward claims of hacking into SNSs accounts. In any event, the spill-over effects of these general cybercrime trends are already affecting SNSs. While the underground pricing lists vary over SNSs hacked accounts that can then be used for other purposes (including many of the activities represented in the categories discussed above like exploitation, direct threats, etc.), an indicative price that was found for a hacked account for SNSs was around $100. The blending of hacking and online harassment creates an additional dynamic for how the spectrum of the dark side of SNSs widens and what that means for victims; however, this has received little research attention (Wilsem, 2013).

11) In the case of gore-sharing activities, a decision has been made to include this category separately as it constitutes a user community that exchanges and amasses ultra-dark content. In bestgore.com, for example, users participate in a community that uploads extremely graphic videos and images including beheadings, executions, impalements, and other events of similar scope. These are being shared and commented upon under claims that blunt reality must be encountered, but at the same time, the legal implications of sharing such content can be brought into question, while issues of users desensitising themselves can surface. Despite the ultra-dark content, there does seem to be research reporting that users exhibit a macabre sense of humour, seek out to criticise systems of power that perpetuate suffering and provide support to troubled users (Alvarez, 2017). However, a prejudicial attitude towards minority groups is also found whose 'lives are deemed less livable and grievable' (Alvarez, 2017, p. 2).

DISCUSSION: TOWARDS COMBINED ELEMENTS AND A FRAMEWORK FOR THE DARK AND THE ULTRA-DARK SIDES

If the combined dark/ultra-dark sides of SNSs is conceived of as a collection of elements then what would these be? Following the emergence of the categories represented above from the empirical data, we now proceed to connect these categories with the elements delineated in Baccarella et al. (2018). Thus, we anchor the ultra-dark side of SNSs-categories that have emerged from the empirical data collection with the elements identified by Baccarella et al. (2018). In this context, we propose either combined categories or explore/describe the structure of those that do not fit the meta-categories, or those that have a more central role to play when considering the development of a framework that captures the dark and the ultra-dark sides. We present the combined categories in Table 2 below, include Figure 3 that illustrates how the categories emerged through visual coding and discuss the implications for these combined categories right after.

Dark honeycomb framework (Baccarella et al.)	**Ultra-Dark elements**	**Combined elements**
Identity (Exploitation of online self)	**Identity**-related activities (Deception, masking, impersonation and de-anonymisation)	Identity exploitation, deception masking and de-anonymisation
Sharing (Inappropriate content and distribution)	**Trading** of child pornography, online child sexual exploitation (**online CSE**) and **gore-sharing** activities	Distributed cyber-criminality (e.g. online CSE) and sharing-based group reinforcement
Groups (in-group and out-group bias)	**Community** reinforcement during sharing	
Presence (Location tracking and monitoring)	**Geolocation** related activities (linked also to de-anonymisation)	Location and co-location exposure and exploitation
Relationships (threat, coercion, abuse and intimidation)	**Direct threats and exploitation** (with possible time-sensitivity)	Direct, indirect and subliminal threats
Reputation (shaming and defamation)	**Subtle exploitation** (exploits and uses reputational risk or results in it)	
Conversations (Misinformation, disinformation and aggressive engagement)	**n/a** (potential link to subliminal threats)	Conversations
n/a	**Procurement-related** activities	Social crime-sourcing
n/a	**Sex**-related activities	Sex-related activities
n/a	**Financial** activities	Financially motivated exploitation

Table 2. A synthesis of the combined elements

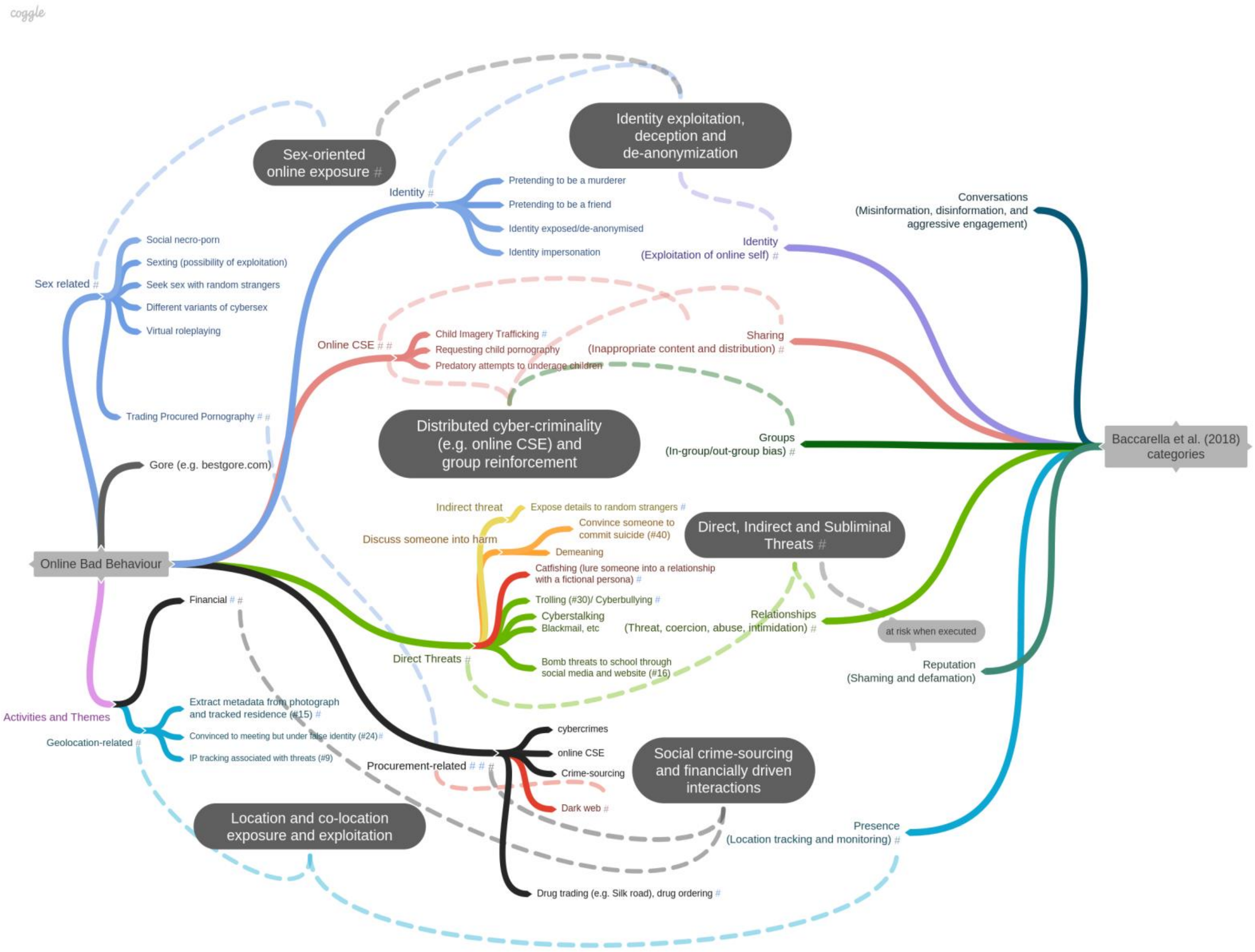

coggle
Sex-oriented online exposure #
Identity exploitation, deception and de-anonymization
Identity #
Pretending to be a murderer
Pretending to be a friend
Identity exposed/de-anonymised
Identity impersonation
Identity (Exploitation of online self) #
Conversations (Misinformation, disinformation, and aggressive engagement)
Sex related #
Social necro-porn
Sexting (possibility of exploitation)
Seek sex with random strangers
Different variants of cybersex
Virtual roleplaying
Trading Procured Pornography # #
Online CSE # #
Child Imagery Trafficking #
Requesting child pornography
Predatory attempts to underage children
Sharing (Inappropriate content and distribution) #
Distributed cyber-criminality (e.g. online CSE) and group reinforcement
Groups (In-group/out-group bias) #
Baccarella et al. (2018) categories
Gore (e.g. bestgore.com)
Indirect threat
Expose details to random strangers #
Discuss someone into harm
Convince someone to commit suicide (#40)
Demeaning
Direct, Indirect and Subliminal Threats #
Catfishing (lure someone into a relationship with a fictional persona) #
Trolling (#30)/ Cyberbullying #
Cyberstalking
Blackmail, etc
Relationships (Threat, coercion, abuse, intimidation) #
at risk when executed
Direct Threats #
Bomb threats to school through social media and website (#16)
Reputation (Shaming and defamation)
Online Bad Behaviour
Financial #
Activities and Themes
Geolocation-related #
Extract metadata from photograph and tracked residence (#15) #
Convinced to meeting but under false identity (#24) #
IP tracking associated with threats (#9)
cybercrimes
online CSE
Crime-sourcing
Social crime-sourcing and financially driven interactions
Procurement-related # # #
Dark web #
Presence (Location tracking and monitoring) #
Location and co-location exposure and exploitation
Drug trading (e.g. Silk road), drug ordering #

Figure 3. Emergent categories of the dark and the ultra-dark through mind-mapping

While the *exploitation of online self* as the dark expression of the identity category is significant (Baccarella et al., 2018), and it does constitute a broad category for describing the victim-related identity-based potential, we find that the concept of identity has a more central role due to its duality. For potential offenders, it is the potential masking of their identity that is of significance. At the same time, for victims, their own identity can be exploited and they can be subjected to deception. This duality must be captured as it signifies a core component of the online space within SNSs. In addition, based on the examples that were discussed previously through the empirical data (e.g. unveiling someone's identity from metadata in photographs), there is also a distinct potential for the de-anonymisation of an online identity that must be considered. This includes situations where a user is trying to protect their online identity but they become compromised. Naturally, the concept of online identity is much more complex and considerably fragmented in an online space (FIDIS, 2009). Nevertheless, we must grant it particular importance because of its centrality.

In their framework, Baccarella et al. (2018) capture sharing (of inappropriate content and distribution) as well as group-related phenomena (with in-group and out-group bias). We confirm the inappropriate content sharing but given that online child sexual exploitation and trading of child pornography is the ultra-dark expression of the sharing category while gore-sharing activities is another one, we combine these elements into distributed cyber-criminality and sharing-based group reinforcement. In these circumstances, trading and sharing activities constitute the mechanism through which groups reinforce their online identities, and continue to carry their activities. Possibly the darkest category discussed, online child exploitation involves one or more of the stages of exploitation in the participation, production and distribution of child pornographic imagery, mostly associated with the dark web. We also find the sharing of horrific and extremely graphic material on bestgore.com or similar networking sites where beheadings, executions, impalements and other acts of similar scope are being shared and commented upon.

On the category related to location, we find that there is almost complete convergence with the Baccarella et al. (2018) framework on *location tracking and monitoring* (presence) where we identify geolocation-related activities as playing a significant role and linked also to the potential for de-anonymisation. Because of the potential of establishing co-location through SNS interactions, which would allow for possible physical harm and exploitation, and because of the potential of tracking users down to their residence or convincing them through deception to go to another location, we combine location exposure and co-location (between offender and victim) exploitation. For this reason, we take location to be another critical characteristic that should be differentiated alongside identity. Like identity, location too has a dual role: offenders have a vested interest for their location to remain shielded (particularly during deception phases) while they try to expose the location of a potential victim. The offender-based protection of location has long been known in cybercrime to constitute a pillar of how cybercriminals handle their own operational security (Butkovic et al., 2019).

In the context of relationships, Baccarella et al. (2018) include *threat, coercion, abuse and intimidation*, while in the context of reputation, they include *shaming and defamation*. We find that there is an important differentiation of threats into *direct* threats (from offender to victim), *indirect* threats (where third parties receive compromising information that they may then act upon) and *subliminal* threats (e.g. convince someone to commit suicide). Perspectives might vary on how the spectrum of threat and abuse is shaped, but overall, threats are also considered to be abusive behaviour (Whittle et al., 2013), while the boundary between such concepts can be considered fluid as the dynamics between offenders and victims are unique and varied. Still, from cyberstalking, to blackmail, trolling, cyberbullying, catfishing, demeaning behaviour, convincing someone to commit suicide or exposing details to random strangers, it is important to recognise that threats have a spectrum and they can be direct (articulation of a threat to a specific person), indirect (exposing details to a third party) and subliminal (a covert threat under a deception strategy like befriending someone while orienting them to harm) while different forms of abusive behaviour are at play. However, through our cases, we perceive reputational risk as a lever that can be used in the context of threats (or abuse, intimidation, etc.). While for some threats, reputation risks (like shaming and defaming) are important, they are perceived here as one of the mechanisms for exercising ongoing pressure during threats or as a consequence when threats are actually operationalised and executed; this contingency can increase the likelihood of the threat being successful or the likelihood that the potential victim will indeed comply with the demands of the offender. Naturally, reputational risks will vary from one user to another and is deemed to be contingent upon the user's own network characteristics and interconnectedness.

Reputational risk is anchored onto the individual exposure that victims have within their own social networks. Some users might not react to a threat that is associated with a reputational risk but could react to other threats when these are associated with financial risk, potential for physical harm, etc.

Additional elements proposed by Baccarella et al. (2018) include *misinformation, disinformation and aggressive engagement* in the context of their category of conversations. While aggressive engagement and disinformation are both identified in the context of deception and deploying subliminal threats, misinformation (in the sense of unintentional distribution) is not observed (possibly due to the nature of the methodology deployed in this research where a question was put directly to the users and the intentionality suggested in the framing of the question). While the empirical data on the ultra-dark had not pointed to a category of conversations *per se*, conversations are perceived as critical and cutting across all other elements. For example, it is through conversations that direct/indirect/subliminal threats are expressed and conducted; it is through conversations that location can be exploited/inferred, or identity deception and de-anonymisation can occur; similarly, it is through conversations that online users find each other and participate in echo chambers that form in-group and out-group biases; it is also through conversations that offenders identify each other and trade child pornographic material or users of gore-sharing communities comment on the material they share. Overall, conversations in SNS are seen here as the main avenue upon which all online SNSs phenomena are constructed. In this context, we find it useful to distinguish conversations between offenders and potential victims from conversations between offenders. Of course, we also have regular online conversations. These shape how online users perceive their own identity online and ultimately how reputational risk might be considered by potential victims if they find themselves in such a position.

In addition, procurement-related activities are considered significant, which we capture through social crime-sourcing and sex-related activities that include the potential for exploitation, which are discussed in the preceding section. Similarly, financial activities are included as a distinct category of potential exploitation, though attempting to defraud users can be done by using a variety of different techniques. The ideas described in this section are synthesised in the framework below and they are reflected upon right after.

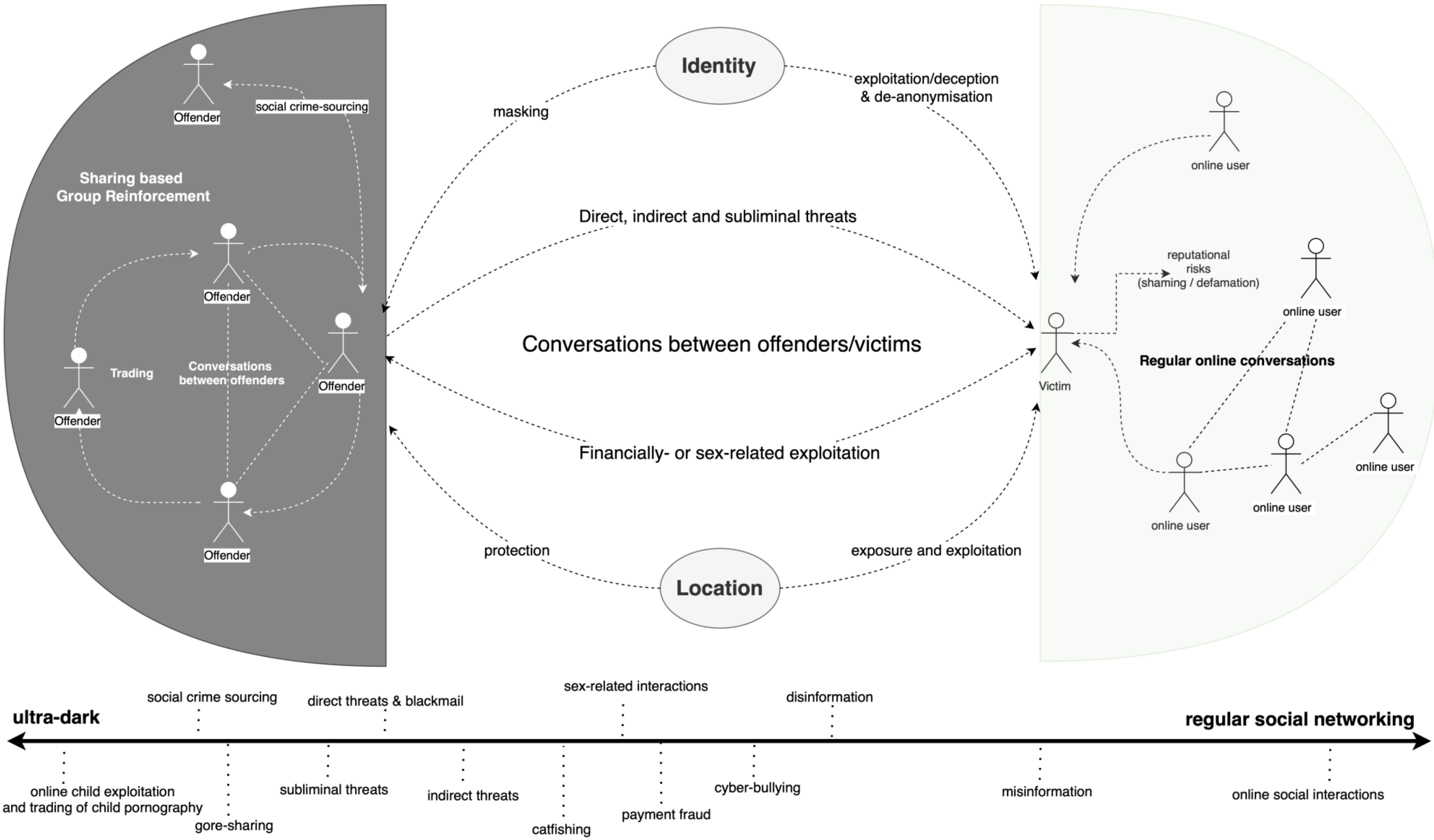


Figure 3. A framework for the dark and ultra-dark sides of SNSs: categories, interactions and the spectrum of exploitation

MANAGING THE RESPONSE TO THE DARK SIDE: MANAGEMENT IMPLICATIONS

From the discussion and the synthesis above that moves from a summary of the categories uncovered through the empirical data collection (Table 1) to the synthesis of the combined elements with the Baccarella et al. (2018) framework (Table 2) and the illustration of the framework in Figure 3, we can see that the phenomena we are exploring in SNSs emerge from a multiplicity of associations between the dark side and the space where regular online conversations take place. Between the two, we consider identity and location as playing key dual roles in the space between the offenders and victims. In the case of identity, offenders can mask their own identity while attempting to exploit the identity of victims, deceive them or deanonymise their identity where possible. In the case of location, offenders try to shield their own location while exploiting/exposing those of their victims, or exploring the possibility for co-location-based exploitation that might lead to physical harm and other serious crimes. Conversations constitute the necessary background in all cases, and we have conversations between offenders and trading on the left-hand side of the diagram and regular online conversations between online users on the right-hand side. In the space between the two, we have conversations between offenders and victims through which direct, indirect and subliminal threats can be manifested, or through which financially motivated exploitation or sex-related exploitation can take place; the latter two can be combined in the case of romance fraud (Whitty, 2018).

Through the proposed framework, we can see that the spectrum of SNS interactions stretches from regular social networking to the ultra-dark side of SNSs; this spectrum cuts across all conversation spaces, creating combinatory possibilities for exploitation. For example, offenders might use the lever of reputational risk that would affect online users in their regular social networking space of interactions, while they conduct social crime-sourcing to take over the account of victims. Thus, exploitation can occur through combinations of vulnerabilities and possibilities. Across the spectrum shown in Figure 3 that stretches from ordinary phenomena to the ultra-dark, we place some indicative phenomena and observe that as we move closer to the ultra-dark side of SNSs, activities are mostly organised around sharing (e.g. gore-sharing) and trading (e.g. child pornography). Through those we have sharing-based group reinforcement (echo-chamber effects) and the emergence of distributed cyber-criminality where the possibilities for social crime-sourcing emerge (cybercrime as a service). Online ultra-dark communities (like traders of child pornographic material) emerge then through the multiplicity of interactions afforded by different SNSs and the dark web. Of course, while the spaces of offenders/victims appear distinct in the diagram (for visual purposes), we must recognise that the very function of all SNSs and the interconnectivity afforded by the internet allows considerably dense interconnections. Thus, while the demarcation between an offender space and a potential victim space is done for analytical purposes, it is reasonable to assume that an online user that can act as an offender may be already part of a (future) victim's online network, and constitute a type of insider threat for the user. Finally, all interactions and conversations will be structured differently in different social networks as each SNS creates different possibilities and path dependencies.

Another important reflection here relates to the spectrum itself and how that is shaped based on different phenomena. As mentioned above, the phenomena we place are indicative and the level of interference of each one, as well as its societal consequences, depends on whether they are considered at the micro-level or at the macro-level. For example, disinformation that is seen as an expression of the dark side of conversations by Baccarella et al. (2018), constitutes the deliberate spread of false information. If conducted between two people (one deliberately sending the false information and another receiving it), it could be conceived of as a gateway to realise other threats (e.g. elicit a reaction from a victim, the output of which can be used to conduct a direct threat or blackmail). However, if disinformation is conducted at the macro-level and *en masse,* for example for political interference in election campaigns, then it is being pulled more into the darker end of the spectrum. The consequences for each phenomenon will vary according to its spread, potency, victim reaction, and a multiplicity of other characteristics that demand further research.

The considerations presented above, allow us to express some key reflections on how the management of the response to the dark side of SNSs should be organised, and how the ultra-dark end of the spectrum of exploitation should be considered for 'dampening' its effects. The management challenges identified need to be tackled together and not in isolation.

First, we need to be paying much closer attention to how SNS interactions are shaped and alert users how the data that they provide during an interaction can be combined with other online data to create exploitation techniques. For instance, given the centrality of identity and location, a social networking platform can assemble proxies/characteristics/attributes that can be used to filter, detect and protect users when their identity or location might be exposed. For example, if a user messages their address, they might be prompted to confirm and verify that they are indeed willing to share their location while alerting to the possibility of geolocation-based exploitation or its combination with other categories. Similarly, dates of birth that constitute a vector of identity, are routinely used for identity fraud or applying for loans while they tend to be largely available in SNSs (Geradts, 2013). Identity is complex but considering some key vectors that could lead to the exposure and subsequent exploitation must be prioritised.

Second, disrupting the stabilisation of new phenomena that gain momentum (e.g. catfishing, cyberstalking, payment fraud through extortion, etc.) needs to be reflected upon. How users react to new phenomena is critical to their stabilisation/destabilisation. Early warning systems of new phenomena of exploitation will be helpful to users. There, the role of SNSs in alerting users about latest exploitation phenomena and activities within each phenomenon is critical in disruption. While the business model of SNSs rests mostly on targeting users for marketing purposes based on the data users themselves make available (demographics, preferences, likes, etc.), the availability of such data can allow SNSs to ‘target their own users’ for safeguarding purposes. This targeting can be customised based on user attributes (e.g. age and gender), and thus, it can acquire a risk-based character. For example, catfishing that aims to solicit funds and seems to be targeting the elderly more and more (also known as romance fraud or sweetheart scam) can be disrupted if SNSs target these users through demographic data they already own and provide support, information and clear reporting lines of help or concerns. Designing an easily accessible and personalised security dashboard with sensible advice for each user could have a considerable impact, as it would allow potential victims to gain heightened visibility of the threat spectrum in SNSs. This can include personalised recommendations on what to do to ameliorate ongoing risks or how to seek help if needed.

Third, in the context of managing the contingencies discussed above, there is a need for SNSs to enhance capabilities to prevent inputs that could lead to exploitation (e.g. geolocation tracking, personal data and interests that could be misused). Also, there is a need to gain capabilities to detect ongoing outputs (e.g. blackmail), and to support users in handling such events. As many users will tend *not* to report such events (Casey, 2004), SNSs must establish more proactive mechanisms for supporting their users and liaising with cybercrime police. Attacking the problem on both ends would allow SNSs to: a) dampen the phenomena of the dark side as fewer cases would materialise, b) disrupt ongoing exploitation (like blackmail) and provide more support to users.

Fourth, through the empirical data we find that the ultra-dark side is more closely expressed through sharing, trafficking and trading activities. As these are important expressions for the ultra-dark side, we suggest that the development of countermeasures must be directed towards restricting trading/trafficking. This should take precedence as this is much more likely to have a wider impact on a series of offences. The disruption of crime-sourcing markets is also an important tenet for the management of these negative effects. Thus, it is imperative that cybercrime police in tandem with SNSs should enhance their capabilities in disrupting trading and trafficking. Also, regulation can be strengthened to increase penalties for those engaged in trading/trafficking activities, and the legality of content-sharing in some gore-sharing websites must be explored further.

Fifth, while SNSs, other technology companies, cybercrime police and local/national governments, constitute a few of the stakeholders that are affected by such phenomena, we need to recognise that this is a multi-stakeholder problem that has a wide reach. Overall, managing the response to the dark and the ultra-dark sides of SNSs implies a negotiation between these stakeholders, and a more comprehensive deconstruction of stakeholder interactions instead of an isolated treatment from each stakeholder. For example, there appears to be a need to create roles for online social workers and SNSs-specialist support groups that can provide support to users. Governments and regulatory bodies need to look into how SNSs can be prompted into action and what meaningful cross-stakeholder collaborations can be established. Without a diverse stakeholder orientation for managing the response to the dark side, an effective handling is unlikely.

Finally, the streamlining of online communication through SNSs creates ideal conditions for the flourishing of deception, which is the cornerstone of subliminal threats, and it can be the road ahead for offenders in the operationalisation of direct and indirect threats, financial exploitation, etc. To enhance user response to the understanding of such conditions, SNSs must manage user perceptions actively and create mechanisms for them. An obvious training approach here would be to allow users to participate in simulated online attacks for different phenomena. This would enhance the resilience of users against potential real attacks and create a 'cyber-vaccination' strategy by exposing users to controlled environments and (ultra-) dark phenomena. It would also allow us to move from a state of having *users* to having *trained users*, who can realise more fully how they might be exploited online. Much like Google has phishing simulations that users can try out, SNSs can create simulations for a variety of dark side phenomena that will enhance user understanding and allow users to practice and realise responses that would protect them. Overall, the boundary between user understanding and deception must be further explored, and prevention must be orchestrated around more comprehensive strategies.

CONCLUSION

Without a doubt, SNSs constitute an important part of modern online life that has allowed billions of people to interconnect, maintain communication between them, as well as share content. However, at the same time, this has led to the emergence of a darker side where exploitation in various forms thrives. The dark side, however, should not be conceived of with a sense of unity. It includes a series of complex interacting phenomena that are dynamic and an ultra-dark side where behaviours range from the hideous to the criminal.

It is in this context that this essay synthesises the key categories of the dark side with the ultra-dark, and provides a framework for conceptualising the categories identified alongside a corresponding spectrum. This provides a more comprehensive architecture of the characteristics of the darker sides of SNSs, and raises the issue of how such phenomena become shaped by variable technological contingencies (e.g. platform-specific characteristics that structure online communication in a particular way).

Managing the response to the dark side implies a reflection of how countermeasures can be developed around disrupting such characteristics. Of course, much more research is required to explore the structure of the darker sides of SNSs, and to explore how interactions between different phenomena give rise to new ones. Also, how transitions occur from the micro-level of (ultra-) dark SNS phenomena to the macro-level is of considerable interest; similarly, organisations can be conceived of as users of SNSs that experience organisational variants of dark-side phenomena from those identified above. In this context, the space between personal SNS use and organisational SNS use requires further exploration for exploring the expression of the dark side. This essay calls researchers to explore the development of the proposed categories further and also to seek complementary or alternative approaches and strategies that will allow us to deconstruct the darker sides of SNSs further. Given the significance of the topic and the complexity and variety of SNSs, scholars must shed much more light on the darker sides of these online phenomena.

REFERENCES

Alvarez, M. (2017). Online spectatorship of death and dying: Pleasure, purpose and community in BestGore.com. *Participations - Journal of Audience and Reception Studies*, *14*(1).

Anderson, R., Barton, C., Böhme, R., Clayton, R., van Eeten, M. J. G., Levi, M., Savage, S. (2013). Measuring the Cost of Cybercrime. In *The Economics of Information Security and Privacy* (pp. 265–300). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-39498-0_12

Angell, I., & Demetis, D. (2010). *Science's first mistake : delusions in pursuit of theory*. London: Bloomsbury Academic. http://dx.doi.org/10.5040/9781472544957.0006

Baccarella, C. V., Wagner, T. F., Kietzmann, J. H., & McCarthy, I. P. (2018). Social media? It's serious! Understanding the dark side of social media. *European Management Journal*, *36*(4), 431–438. https://doi.org/10.1016/j.emj.2018.07.002

Barton, H. (2016). *The dark side of the Internet*. *An introduction to cyberpsychology*. London:

Routledge. https://doi.org/10.4324/9781315741895

Bluth, K., & Blanton, P. W. (2015). The influence of self-compassion on emotional well-being among early and older adolescent males and females. *The Journal of Positive Psychology, 10*(3), 219–230. https://doi.org/10.1080/17439760.2014.936967

Butkovic, A., Mrdovic, S., Uludag, S., & Tanovic, A. (2019). Geographic profiling for serial cybercrime investigation. *Digital Investigation*, *28*, 176–182. https://doi.org/10.1016/j.diin.2018.12.001

Capra, F., & Luisi, P. L. (2014). *The Systems View of Life*. *The Turning Point: Science, Society, and the Rising Culture*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511895555

Casey, E. (2004). Reporting security breaches – a risk to be avoided or responsibility to be embraced? *Digital Investigation*, *1*(3), 159–161. https://doi.org/10.1016/j.diin.2004.07.008

Chou, C. H., Sinha, A. P., & Zhao, H. (2008). A text mining approach to Internet abuse detection. *Information Systems and E-Business Management*, *6*(4), 419–439. https://doi.org/10.1007/s10257-007-0070-0

Christopherson, K. M. (2007). The positive and negative implications of anonymity in Internet social interactions: "On the Internet, Nobody Knows You're a Dog." *Computers in Human Behavior*, *23*(6), 3038–3056. https://doi.org/10.1016/j.chb.2006.09.001

Clement, J. (2019). Number of monthly active Facebook users worldwide as of 3rd quarter 2019. Retrieved from https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/

Craker, N., & March, E. (2016). The dark side of Facebook®: The Dark Tetrad, negative social potency, and trolling behaviours. *Personality and Individual Differences*, *102*, 79–84. https://doi.org/10.1016/j.paid.2016.06.043

Davis, M., & Whalen, P. J. (2001). The amygdala: vigilance and emotion. *Molecular Psychiatry, 6*(1), 13–34. https://doi.org/10.1038/sj.mp.4000812

Demetis, D.S. (2018a). How the 'Original Internet Godfather' walked away from his cybercrime past – interview. Retrieved from https://theconversation.com/how-the-original-internet-godfather-walked-away-from-his-cybercrime-past-interview-88822

Demetis, D. S. (2018b). Fighting money laundering with technology: A case study of Bank X in the UK. *Decision Support Systems*, *105*, 96–107. https://doi.org/10.1016/j.dss.2017.11.005

Dhillon, G., & Backhouse, J. (2001). Current directions in IS security research: towards socio-organizational perspectives. *Information Systems Journal*, *11*(2), 127–153. https://doi.org/10.1046/j.1365-2575.2001.00099.x

FIDIS. (2009). D17.1 Modelling New Forms of Identities: Applicability of the Model Based on Virtual Persons . Retrieved from http://www.fidis.net/fileadmin/fidis/deliverables/fidis-wp17-del17.1.Modelling_New_Forms_of_Identities.pdf

Fox, J., & Moreland, J. J. (2015). The dark side of social networking sites: An exploration of the relational and psychological stressors associated with Facebook use and affordances. *Computers in Human Behavior*, *45*, 168–176. https://doi.org/10.1016/j.chb.2014.11.083

Furnell, S. (2003). Cybercrime: Vandalizing the information society. *Web Engineering, Proceedings*. https://doi.org/10.1007/3-540-45068-8_2

Geradts, Z. (2013). Identity Theft. In *Encyclopedia of Forensic Sciences: Second Edition*. https://doi.org/10.1016/B978-0-12-382165-2.00217-8

Goldsworthy, T., Raj, M., & Crowley, J. (2017). "Revenge porn": An analysis of legislative and policy responses. *International Journal of Technoethics*, *8*(2), 26–41. https://doi.org/http://dx.doi.org/10.4018/IJT.2017070103

Isenberg, P., Elmqvist, N., Scholtz, J., Cernea, D., Ma, K.-L., & Hagen, H. (2011). Collaborative visualization: definition, challenges, and research agenda. *Information Visualization*, *10*(4), 310–326. https://doi.org/10.1177/1473871611412817.

Johnson, N. (2019). The dark side of social media. *Physics World*, *32*(3), 32–36. https://doi.org/10.1088/2058-7058/32/3/31

Kietzmann, J. H., Hermkens, K., McCarthy, I. P., & Silvestre, B. S. (2011). Social media? Get serious! Understanding the functional building blocks of social media. *Business Horizons*, *54*(3), 241–251. https://doi.org/10.1016/j.bushor.2011.01.005

Kircaburun, K., & Griffiths, M. D. (2018). The dark side of internet: Preliminary evidence for the associations of dark personality traits with specific online activities and problematic internet use. *Journal of Behavioral Addictions*, *7*(4), 993–1003. https://doi.org/10.1556/2006.7.2018.109

Klein, H., & Myers, M. (1999). A set of principles for conducting and evaluating interpretive field studies in information systems. *MIS Quarterly*, *23*(1), 67–94.

Křemen, P., Mička, P., Šmíd, M., & Blaško, M. (2012). Ontology-driven mindmapping. In *ACM*

International Conference Proceeding Series.

Lee, A. S., & Baskerville, R. L. (2003). Generalizing Generalizability in Information Systems Research. *Information Systems Research*, *14*(3), 221–243. https://doi.org/10.1287/isre.14.3.221.16560

Lee, Y. K., Chang, C.-T., Lin, Y., & Cheng, Z.-H. (2014). The dark side of smartphone usage: Psychological traits, compulsive behavior and technostress. *Computers in Human Behavior*, *31*(1), 373–383. https://doi.org/10.1016/j.chb.2013.10.047

Lefever, S., Dal, M., & Matthíasdóttir, Á. (2007). Online data collection in academic research: Advantages and limitations. *British Journal of Educational Technology*, *38*(4), 574–582. https://doi.org/10.1111/j.1467-8535.2006.00638.x

Letichevsky, A. A., Letychevskyi, O. O., Skobelev, V. G., & Volkov, V. A. (2017). Cyber-Physical Systems. *Cybernetics and Systems Analysis*, *53*(6), 821–834. https://doi.org/10.1007/s10559-017-9984-9

Lilley, P. (2000). *Dirty Dealing*. London: Kogan Page.

Mazza, R. (2009). *Introduction to information visualization*. London: Springer.

McMurdie, C. (2018). Paedophile rings “one step ahead of cyber criminals” claims former head of NCCU. Retrieved from https://tech.newstatesman.com/security/paedophile-cyber-crime

Omotoyinbo, F. R. (2014). Online Radicalisation: The Net or the Netizen? *Social Technologies*, *4*(1), 51–61. https://doi.org/10.13165/ST-14-4-1-04

Roshan, M., Warren, M., & Carr, R. (2016). Understanding the use of social media by organisations for crisis communication. *Computers in Human Behavior*, *63*, 350–361. https://doi.org/10.1016/j.chb.2016.05.016

Ross, M. W. (2005). Typing, doing, and being: Sexuality and the internet. *Journal of Sex Research*, *42*(4), 342–352. https://doi.org/10.1080/00224490509552290

Ruiz Vicente, C., Freni, D., Bettini, C., & Jensen, C. S. (2011). Location-related privacy in Geo-social networks. *IEEE Internet Computing*, *15*(3), 20–27. https://doi.org/10.1109/MIC.2011.29

Salo, J., Mäntymäki, M., & Islam, A. K. M. N. (2018). The dark side of social media – and Fifty Shades of Grey introduction to the special issue: the dark side of social media. *Internet Research*, *28*(5), 1166–1168. https://doi.org/10.1108/IntR-10-2018-442

Shams, L., & Seitz, A. R. (2008). Benefits of multisensory learning. *Trends in Cognitive*

Sciences, *12*(11), 411–417. https://doi.org/10.1016/j.tics.2008.07.006

Shelton, J., Eakin, J., Hoffer, T., Muirhead, Y., & Owens, J. (2016). Online child sexual exploitation: An investigative analysis of offender characteristics and offending behavior. *Aggression and Violent Behavior*, *30*, 15–23. https://doi.org/10.1016/j.avb.2016.07.002

Silic, M., & Back, A. (2016). The dark side of social networking sites:Understanding phishing risks. *Computers in Human Behavior*, *60*, 35–43. https://doi.org/10.1016/j.chb.2016.02.050

Spangler, S., Kreulen, J. T., & Lessler, J. (2002). MindMap: utilizing multiple taxonomies and visualization to understand a document collection. In *Proceedings of the 35th Annual Hawaii International Conference on System Sciences* (pp. 1170–1179). IEEE Comput. Soc. https://doi.org/10.1109/HICSS.2002.994039

Spence, R. (2007). *Information visualization : design for interaction* (2nd ed.). New York: Addison Wesley.

Spitzberg, B. H., & Hoobler, G. (2002). Cyberstalking and the technologies of interpersonal terrorism. *New Media and Society*, *4*(1), 71–92. https://doi.org/10.1177/14614440222226271

Spitzer, M., Fischbacher, U., Herrnberger, B., Grön, G., & Fehr, E. (2007). The Neural Signature of Social Norm Compliance. *Neuron*, *56*(1), 185–196. https://doi.org/10.1016/j.neuron.2007.09.011

Stroud, S. R. (2014). The Dark Side of the Online Self: A Pragmatist Critique of the Growing Plague of Revenge Porn. *Journal of Mass Media Ethics: Exploring Questions of Media Morality*, *29*(3), 168–183. https://doi.org/10.1080/08900523.2014.917976

Synnott, J., Coulias, A., & Ioannou, M. (2017). Online trolling: The case of Madeleine McCann. *Computers in Human Behavior*, *71*, 70–78. https://doi.org/10.1016/j.chb.2017.01.053

van Zoonen, W., Verhoeven, J. W. M., & Vliegenthart, R. (2016). Social media's dark side: inducing boundary conflicts. *Journal of Managerial Psychology*, *31*(8), 1297–1311. https://doi.org/10.1108/JMP-10-2015-0388

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*, 1146–1151. https://doi.org/10.1126/science.aap9559

Walsham, G. (1995). The Emergence of Interpretivism in IS Research. *Information Systems Research*, *6*(4), 376–394. https://doi.org/10.1287/isre.6.4.376

Walters, J. (2011). *Anti-money laundering and counter-terrorism financing across the globe : a* *comparative study of regulatory action*. *AIC reports Research and public policy series,*. Canberra: Australian Institute of Criminology. Retrieved from http://www.aic.gov.au/publications/current series/rpp/100-120/rpp113.aspx

Whittle, H., Hamilton-Giachritsis, C., Beech, A., & Collings, G. (2013). A review of online grooming: Characteristics and concerns. *Aggression and Violent Behavior*, *18*(1), 62–70. https://doi.org/10.1016/j.avb.2012.09.003

Whitty, M. T. (2018). Do You Love Me? Psychological Characteristics of Romance Scam Victims. *Cyberpsychology, Behavior, and Social Networking*, *21*(2), 105–109. https://doi.org/10.1089/cyber.2016.0729

Wilsem, J. van. (2013). Hacking and Harassment—Do They Have Something in Common? Comparing Risk Factors for Online Victimization. *Journal of Contemporary Criminal Justice*, *29*(4), 437–453. https://doi.org/10.1177/1043986213507402

Zhang, X.-Z., Liu, J.-J., & Xu, Z.-W. (2015). Tencent and Facebook Data Validate Metcalfe’s Law. *Journal of Computer Science and Technology*, *30*(2), 246–251. https://doi.org/10.1007/s11390-015-1518-1