Social media as a disruptive social innovation : new possibilities for moral enforcement and the emergence of a Fourth Wave of feminism

David BERTRAND,
doctoral student,
Université de Bordeaux

Social media as a disruptive social innovation : new possibilities for moral enforcement and the emergence of a Fourth Wave of feminism

The present paper will expose the hypothesis that social media encourage the development of a very efficient kind of social control, fueling the creation of social norms by political activists. I will use the example of online feminism and its repression of sexist speech, mostly in France, since I submitted a survey on feminist Facebook pages and Tumblrs.

I will draw on Boyd and Ellison's definition of social network sites as sites that allow users to (1) create a « public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system » (Boyd & Ellison, 2008). Nonetheless, even if these characteristics are shared by all social network sites, the websites differ on the ways their architectures organize visibility, invisibility, the flow and direction of information, and the shape of social networks : drawing on these points, **the regime of visibility built on Facebook (but also Twitter) is a panopticon, where everyone can (and is incited to) monitor everyone else** ; **and where everyone knows that (almost) anyone else can see what they do** (Bucher, 2014).

Social norms are often defined as **« shared understandings about actions that are obligatory, permitted, or forbidden** (Crawford and Ostrom, 1995). **»** (Ostrom, 2000). However, it is not possible to prove the existence of a norm by simply observing a regular behavior (or the absence of it), nor by the existence of the beliefs and values underlying it. That's why our operative definition of a social norm must include **the punishment of those who don't follow it**, following Robert Axelrod's evolutionary approach to norms (Axelrod, 1986). **Therefore, it is possible to demonstrate that there is a social norm in a social setting when a particular behavior (or its absence) tends to be punished by agents**. We

can then identify cooperators (who simply obey the norm), violators (who don't obey) and punishers (who punish violators).

The panoptic architecture of social network sites such as Facebook and Twitter is of much utility here : it **allows social scientists to observe the enforcement of social norms by witnessing punishments**, arguments and disputes over speeches and behaviors. Moreover, due to the searchability and persistence of information on the Internet, scientists can find and observe almost any interaction that hasn't been submitted to filters or on private channels. Of course, this implies for scientists to be careful about the ethics of privacy and to respect the context of speech and the intentions of the users (Latzko-Toth & Proulx, 2013).

In our specific case, we will adress **the creation and enforcement, by feminist activists, of social norms against sexist speech on social media**. Indeed, the practice of public shaming by feminist activists against sexism on social media is a well-known and publicised fact (Munro, 2013). The most mediatized case involved Nobel Prize Tim Hunt's allegedly sexist remarks when he made an attempt to tell a bad joke on a conference. Moreover, **sexist language adresses women's social identity and is thus likely to trigger strong emotional reactions in some people**. Contents provoking this kind of strong emotional reactions, and particularly those causing anger (but also awe), are more likely to spread and attract attention on social media (Berger & Milkman, 2011), but also to incite witnesses to exert social control if a social norm exists (Chaurand & Brauer, 2008). Sexist language is also a well discussed topic in feminist philosophy and litterature, as well as sexist speech online and cyberviolence against women.

Thus, I will **use the possibilities offered by social media to observe symbolic punishment against sexist speech on social network sites and try to explain how such a social norm can spread in this environment**. I will first show that game theory models of social norms

provide a good framework to understand public shaming on social network sites, and then confront it to empirical data gathered on an online survey. Finally, we will discuss the practical implications of such findings.

I – Public shaming practices according to game theory models

As it has already been told, the panoptic environment of social network sites allows horizontal surveillance of actions. This enables a particular form of norm enforcement, sometimes called online firestorms, which are **« large amounts of critique, insulting comments, and swearwords against a person, organization, or group […] formed by, and propagated via, thousands of millions of people within hours »** (Rost, Stahel & Frey, 2016). This phenomenon has been related to public shaming in a growing « call-out culture » (Munro, 2013), albeit **it must be stressed that public shaming should be more precisely linked to the specific action of publicising an event and framing it in a way likely to shame the culprit and provoke an online firestorm against him or her**. Public shaming is therefore **a conscious and deliberate act of enforcing a social norm by submitting the violator to the public's judgement**, which is a form of moral punishment linked to **reputation**. It has also been called **vigilantism** (Laidlaw, 2017). The fact that social movements or activists use it as a tool for social change is not very surprising then.

Indeed, the panoptic architecture of social media allows reputation-based social interactions. This type of social setting has been theorized in evolutionary game theory of social norms to be the most likely to support the creation and enforcement of social norms. « **All that matters in these models is that agents can properly identify other agents, such that they can maintain a record of their past behavior. This allows for the possibility of reputations : people who have the reputation of being cooperative will be treated**

cooperatively, and those who have a reputation of being unfair will be treated unfairly » (Axelrod, 1986). Thus, the reputation of an individual among certain groups becomes a valuable asset, beyond the simple fact that a low reputation is likely to isolate an individual from a group or even the community as a whole. **Moral judgements are, in this situation, « a powerful tool for maintaining social order and motivating actions that benefit groups or group members »** (Simpson, Willer & Harrell, 2017). Moral judgements affect the reputation of the individuals, and help identifying them as violators or cooperators : that's why their visibility is a key element to enforce social norms. Moreover, we must acknowledge that, besides the material cost of a bad reputation, **humans are intrinsically motivated to be seen as high-morality members of their community, they have a predisposition to obey social norms (Greene, 2017).**. High visibility and a **propension to morality** can explain how social groups can overcome the second-order public good dilemma (why should an individual punish the violation of a norm if such punishment has a cost ?). « Thus, **when people are able to convey norms and monitor each other's behavior through reputation spreading, costly punishment is less needed**, unless gossip turns out to be ineffective » (Wu, Balliet & Van Lange, 2016**). This means that cooperation is better achieved through gossip and reputation-spreading than through material punishment** (Wu, Balliet & Lange, 2016). For example, Elinor Ostrom's model on the evolution of social norms stresses that **complete information about the types of players (cooperators or violators) strongly favours cooperators to a norm – to the point of making violators disappear** (Ostrom, 2000). Of course, even with a panoptic regime of visibility, we must acknowledge that a state of complete information is unlikely to occur. But « **if there's a noisy signal about a player's type that is at least more accurate than random, trustworthy types will survive as a substantial proportion of the population** » (Ostrom, 2000). Such signals exist on social media, and are very efficient. **Facebook pages can be created by moral entrepreneurs**

(intrinsically motivated individuals) to gather cooperators or even punishers, who are motivated to execute a punishment against violators of the norm they support. We can take here the example of french Facebook page « L'empêcheuse de penser en rond »[1], which goal is to find sexist publications, massively report them in order to obtain their banishment, and then to publish the action on the page. In this context, violators are publicly shamed and punishers can be identified. On Twitter, feminists can signal themselves in their profile (picture, description) and their publications, can mutually follow each other and cooperate on specific *hashtags*. The political polarization on Twitter may be one consequence of these features, with clusters of like-minded individuals retweeting each other and mentionning ideological opponents to provoke confrontation (Conover et al. , 2011). This kind of cooperative monitoring and reporting is a very good example of what can be acheived in an environment allowing defenders of a social norm to signal themselves, bond together and coordinate their actions.

We must add that, **« As people enforce social norms and promote public goods, it is most likely that they perceive the behavior of the accused public actors as driven by lower-order moral ideals and principles while that they perceive their own behavior as driven by higher-order moral ideals and principles. From this point of view there is no need to hide their identity** » (Rost, Stahel & Frey, 2016). Punishment has then no cost and can even be a payoff by itself for some members of a social group.

Others have used cost and benefits analysis and empirical data to enlighten the potential of online environments in fostering norm enforcement, and we will quote them to summarize. **« Specifically, research shows that Olson's second-order public good dilemma can be overcome if (1) norm enforcement is cheap, i.e., it occurs in low cost situations, (2) additional benefits are provided to the norm enforcers that disproportionately motivate**

[1] The page was created on july 2015 and has almost 93 000 followers on Facebook in september 2017.

them compared to non-enforcers, i.e. selective incentives are present and / or (3) if some individuals are present that are intrinsically motivated to enforce norms, i.e., some amount of altruistic punishment occurs » (Rost, Stahel & Frey, 2016). These conditions are met on several social network sites, and particularly in the case that we are studying. We can see that evolutionary game theory applied to social norms explains well how feminists can enforce social norms against sexism on social network sites.

Moreover, **public shaming, since it is public**, allows other users of the social network site not only to identify who are the violators and the enforcers of a specific norm, but also – and more importantly – to learn which behaviors are punished. This **modifies their normative expectations (what they expect others will think about their actions) ; but also their empirical expectations (what they think others will do)** (Bicchieri & Muldoon, 2014). Thus, they can adapt their behaviour accordingly, be it by conformist transmission or by rational pay-off-based calculation (Heinrich & Boyd, 2001). The greater visibility of actions induced by the panoptic environment enhances the – already strong – power of social network in influencing individual behavior.

II – The use of public shaming by feminist activists online in France

Besides feminist Facebook pages, Twitter accounts or famous stories, we tried to measure the frequency of public shaming practices against sexism by feminists on the internet through a survey. The sample gathers 340 respondents, who answered voluntarily to a survey published in several french feminist Facebook pages and Tumblrs. Of course, the recruitment process skews the results since it only studies practices and attitudes of feminists online, and is therefore not representative of the general population, nor of feminists who don’t use social media. Still, these feminists are an interesting sample to study since they are embedded in social media and follow feminist pages, and thus more likely to fit the *Idealtypus* studied

here : the feminist activist who uses social media in their daily life. The survey shows that : 67% of the sample strongly agrees with forbidding sexist speeches, images or advertisements (table 1) ; feminist identity is strongly correlated with the use of social network sites for activism purposes (though it may seem obvious, it is still important to test such affirmations) (table 2) ; 51% of the sample reported personally practicing public shaming, 54,5% reported using social network sites tools to signal or ban sexist content, 63,6% reported relaying public shaming content, and even 8,4% admitted doxxing individuals they considered sexists (table 3).

Data confirms then that feminists use public shaming, reporting and even doxxing in order to adress sexist speech online. These observations, when added to models of social norm theory, demonstrate that there are indeed groups enforcing social norms against sexism on social network sites, challenging the idea that the Internet is a jungle with no rules.

III – Norm violation online

Nevertheless, even in such conditions, violators of a norm will never cease to exist. They can continue to violate the norm for various reasons.

For example, a typical phenomenon on social media is the existence of the so-called *trolls*, to whom the general public generally attributes the behavior to a genuine interest in bothering others. We will consider it as an inherent threat on social network since it has no implications here.

Another reason for the persistence of violators of social norms despite their spreading and enforcement stems from the fact that **social network sites are diverse and, therefore, various social norms and moral entrepreneurs can coexist, leading to conflicts of norms in particular cases, or even anti-social punishment**. In our case, people calling out violators

of the norm they defend can themselves be punished by members of an outgroup, usually by harassment practices. **However, it does not seem that anti-social punishment is executed at an equal or higher rate than pro-social punishment. Indeed, if feminists are organized and punish a certain pourcentage of violators of the norm against sexist speech (helped by specific tools such as reporting / banning options and moderators), their opponents can surely harass a certain pourcentage of the ones identified as punishers, but can't realistically punish cooperators (the ones who just refrain from committing sexist actions or speeches). Then, the empirical and normative expectations of the general users remain unchanged about what is the right thing to do, and if anti-social retaliation surely can slow down the spread of this social norm and discourage moral entrepreneurs, it probably cannot make it disappear nor replace it by another norm.** From this, it seems that a social norm trying to ban a behaviour (here sexist speech) will tend to have more success than a hypothetical social norm protecting the right to adopt any behaviour (for example a norm encouraging free speech, since a freedom can not be enforced)[2].

The third reason why norm violation can persist in our environment is **error or misinterpretation. The scope of behaviours violating a norm is not always clear**, and there can be competing subnorms about what is the correct meaning of the main norm. In our case, the definition of a sexist act of speech may not be the same for everyone. Some acts of speech also display a certain level of **ambiguity**, such as art or humour, and will therefore more likely be causes of conflict about whether or not the author should be punished. The author herself could be well-intented and still punished, etc. The case of Tim Hunt can be reminded here. **In fact, the more the norm enforcer seems aggressive towards a norm violation perceived as benign or benevolent, the more the observers to be likely to assign**

[2] Nassim Nicholas Taleb comes to a similar conclusion, but from the observation of the existence of asymetrical preferences, see (Taleb, 2016).

a bad reputation to the punisher (Eriksson, Andersson and Strimling, 2017). **The ambiguity or misunderstandings around a social norm on social network sites can lead to involuntary violations, so reactions from the norm enforcers will be perceived as disproportionate ; and we suggest this gap might sometimes be one of the causes of what is taken for anti-social punishment but could be the enforcement of norms about what is acceptable or not in terms of norm enforcement.**

Conclusion : Social norm enforcement online and offline

To conclude, we shall recall that we have demonstrated, drawing on well-known game theory models of social norms, how social network sites provide powerful tools of social control for feminist activists – though the conclusion may be extended to other groups or movements. Social control is encouraged on social media by the **regime of visibility**, the **possibilities for organizing actions and gathering group members**, and the **specific moderation tools provided by social network sites**. **However, vigilantism will not make violation of social norms disappear forever**, since this implies **complete information**, **systematic reaction from vigilantes**, a **perfect and homogeneous understanding of the enforced social norm from all the users**, as well as the **disappearing of the residual troll** (i.e. the individual intrinsically motivated to break social norms) **and of competing social norms**.

We also shall warn that **the aforementioned phenomenons and frameworks take place in a particular time and place, where specific political opportunities are set, and influence the possibilities of social movements**. The fact that campaigns against manspreading are quickly organized on public transportation services in a growing number of cities around the world[3] is a good example : the publicization of this behaviour and its framing as a gender-based offence may have been done on the Internet, but **the adoption of campaigns by**

[3] NewYork, Tokyo, Los Angeles, Seoul, Madrid, Bordeaux, etc.

institutions tied to public services would probably not have happened if there was not a favorable political environment for it.

I suggest that the existence of such a propicious political opportunity structure, encouraging the rise of public debate and reform around gender issues (at least but not only) in France, as well as the new tools provided to activists by social media are some of the features supporting the hypothesis of the rise of a Fourth Wave of feminism.

References

- R. Axelrod, (1986). « An evolutionary approach to norms », *The American Political Science Review*, Vol. 80, n°4, 1986, pp. 1095-1111.
- H. Becker, (1985). *Outsiders : études de sociologie de la déviance*, Paris, Métailié.
- J. Berger, C. Milkman, (2012). « What makes online content viral ? », *Journal of marketing research*, Vol. 49, Issue 2, pp. 192-205.
- G. Bronner, (2013). *La démocratie des crédules*. Paris, Presses universitaires de France.
- C. Bicchieri, E. Xiao, (2009). « Do the right thing : but only if others do so », *Journal of Behavioral Decision Making*, 22, pp. 191-208.
- D. Boyd, N. Ellison (2008). « Social Network Sites : definition, history and scholarship », *Journal of Computer-mediated Communication*, 13, pp. 210-230.
- T. Bucher, (2014). « Want to be on the top? Algorithmic power and the threat of invisibility on Facebook », *New Media & Society*, 14(7), pp. 1164–1180.
- D. Cardon, (2010). *La démocratie Internet : promesses et limites*. Paris, Seuil (Coll. La République des idées).
- N. Chaurand, M. Brauer (2008). « What determines social control ? People's reactions to counternormative behaviors in urban environments », *Journal of Applied Social Psychology*, 38, 7, pp. 1689-1715.
- CONOVER *et al.* , « Political polarization on Twitter », *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*., 2011.
- K. Eriksson, P. Andersson, P. Strimling, (2017). « When is it appropriate to reprimand a norm violation ? The roles of anger, behavioral consequences, violation severity, and social distance », *Judgement and decision making*, Vol. 12, n°4, pp. 396-407.

- J. Greene, (2017). *L'émotion, la raison et tout ce qui nous sépare*. Paris, Lettonie, Editions Markus Haller.
- J. Heinrich, R. Boyd, (2001). « Why people punish defectors : weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas », *Journal of theoretical biology*, n° 208, pp. 79-89.
- E. Laidlaw, (2017). « Online Shaming and the right to privacy », *Laws*, 6, 3. Doi:10.3390/laws6010003.
- G. Latzko-Toth, S. Proulx, (2013) « Enjeux éthiques de la recherche sur le web », in C. Barats, *Manuel d'Analyse du Web*, (pp. 32-52), Paris, Armand Colin.
- E. Munro, (2013). « Feminism : a fourth wave ? », *Political studies association*. URL : https://www.psa.ac.uk/insight-plus/feminism-fourth-wave
- E. Ostrom, (2000). « Collective action and the evolution of social norms » , *The journal of economic perspectives,* Vol. 14, n°3, pp. 137-158.
- K. Rost, L. Stahel, B. Frey, (2016). « Digital social norm enforcement : online firestorms in social media », *PLoS ONE*, 11 (6).
- B. Simpson, R. Willer, A. Harrell, (2017) « The enforcement of moral boundaries promotes cooperation and prosocial behavior in groups », *Nature : scientific reports*, 7 :42844. doi:10.1038/srep42844
- P. Sobkowicz. , A. Sobkowicz , (2010). « Dynamics of hate based Internet user networks », The *European Physical Journal B*, Volume 73, Issue 4, pp 633–643.
- N. Taleb, (2016) « The most intolerant wins : the dictatorship of the small minority », in (incoming) N. Taleb, *Skin in the* game, (2018). Retrieved from *Medium.com*, URL : https://medium.com/incerto/the-most-intolerant-wins-the-dictatorship-of-the-small-minority-3f1f83ce4e15#.r02304p5c.

- J. Wu, D. Balliet, P. Van Lange, (2016). « Gossip versus punishment : the efficiency of reputation to promote and maintain cooperation », *Nature scientific reports*. doi:10.1038/srep23919.

Table 1.

Answers of the sample to the following statement : « Sexist speeches, images and advertisements should be banned ».

	Effectifs	% Obs.
Totally disagree	4	1,2%
Somewhat disagree	15	4,4%
Maybe	24	7,1%
Somewhat agree	78	22,9%
Totally agree	219	64,4%
Total	**340**	**100%**

Table 2.

Use of social media according to strength of feminist identity.

Do you sometimes use social media to discuss or promote feminism ? →	**Yes**		**No**		**Total**	
Do you consider yourself a feminist ? ↓	Eff.	% Obs.	Eff.	% Obs.	**Eff.**	**% Obs.**
Not at all	0	0%	0	0%	**0**	**100%**
Rather not	0	0%	0	0%	**0**	**100%**
It depends	6	46,2%	7	53,8%	**13**	**100%**
Somewhat	58	70,7%	24	29,3%	**82**	**100%**
Totally	222	90,6%	23	9,4%	**245**	**100%**
Total	**286**	**84,1%**	**54**	**15,9%**	**340**	

Percentages are calculated for the observations on lines (not columns)

Table 3.

Strategies and practices of activism online.

	Effectifs	% Obs.
Reporting / suppression of sexist / antifeminist images or publications using the tools available on the social network site	**156**	**54,5%**
Denunciation of sexist / antifeminist comments, images or actions by making them public on a social network site (public shaming)	**146**	**51%**
Research and disclosure of information about the authors of such speeches or actions (doxxing)	**24**	**8,4%**
Diffusion or sharing of publications or tweets denouncing or publicizing sexist actions or remarks on your network (supporting and relaying public shaming)	**182**	**63,6%**
Discussions with other activists or individuals suporting your views	**164**	**57,3%**
Discussions with individuals holding opposing views	**102**	**35,7%**
Content creation	**53**	**18,5%**
Feminist content diffusion (share, retweet, etc.)	**235**	**82,2%**
Organisation of feminist meetings, events or protests	**43**	**15%**
Other	**10**	**3,5%**
Total	**286**	*****

***** : Total percentage is higher than 100% because participants could select multiple choices.