# Digital Wildfires: Propagation, Verification, Regulation and Responsible Innovation

HELENA WEBB, University of Oxford
PETE BURNAP, Cardiff University
ROB PROCTER, University of Warwick
OMER RANA, Cardiff University
BERND CARSTEN STAHL, De Montfort University
MATTHEW WILLIAMS, Cardiff University
WILLIAM HOUSLEY, Cardiff University
ADAM EDWARDS, Cardiff University
MARINA JIROTKA, University of Oxford

Social media platforms provide an increasingly popular means for individuals to share content online. Whilst this produces undoubted societal benefits, the ability for content to be spontaneously posted and re-posted creates an ideal environment for rumour and false/malicious information to spread rapidly. When this occurs it can cause significant harms and can be characterised as a 'digital wildfire'. In this paper we demonstrate that the propagation and regulation of digital wildfires form important topics for research and conduct an overview of existing work in this area. We outline the relevance of a range of work from the computational and social sciences, including a series of insights into the propagation of rumour and false/malicious information. We argue that significant research gaps remain - for instance there is an absence of systematic studies on the effects of digital wildfires and there is a need to combine empirical research with a consideration of how the responsible governance of social media can be determined. We propose an agenda for research that establishes a methodology to explore in full the propagation and regulation of unverified content on social media. This agenda promotes high quality interdisciplinary research that will also inform policy debates.

Categories and Subject Descriptors: **K.4.1 [Public Policy Issues]**: Abuse and crime involving computers, ethics, regulation

General Terms: Human Factors, Measurement

Additional Key Words and Phrases: social media, rumour,

**ACM Reference Format**:

Helena Webb, Pete Burnap, Rob Procter, Omer Rana, Bernd Carsten Stahl, Matthew Williams, William Housley, Adam Edwards and Marina Jirotka. 2016. Digital Wildfires? Propagation, Verification, Regulation and Responsible Innovation. *ACM Trans on Info Systems*. Special issue on Trust and Veracity of Information in Social Media. Number Article number (date) number of pages.
DOI:http://dx.doi.org/10.1145/0000000.0000000

This work is supported by the Economic and Social Research Council, grant reference ES/L013398/1

First Author address: Helena Webb, Department of Computer Science, Wolfson Building, Parks Road, University of Oxford, Oxford, OX1 3QD, United Kingdom.

1046-8188/2015/MonthOfPublication - ArticleNumber $15.00
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
DOI:http://dx.doi.org/10.1145/0000000.0000000

## 1. INTRODUCTION: SOCIAL MEDIA, UNVERIFIED CONTENT AND DIGITAL WILDFIRES

Social media platforms such as Twitter, Facebook, Instagram and Tumblr etc. provide a highly accessible and increasingly popular means for individuals to share content. In addition to producing their own posts, users of these platforms are able - and encouraged - to share, forward or retweet posts made by others. The combination of mass uptake and instant propagation enables user-generated content to spread at a rate that is exponentially faster than traditional 'word of mouth' [Murthy 2012a; Edwards et al. 2013]. Social media platforms thus serve to accelerate and amplify communicative acts. This acceleration and amplification can in turn produce considerable societal impacts. Beneficial impacts include enhanced social resilience in the face of natural disasters (for example, through the spread of solidarity messages), the democratisation of news media and the facilitation of widening participation in civil society [Mossberger 2008 et al.; Loader and Mercea 2012; Dahlgren 2014]. However, social media is also observed to produce harmful impacts. The ability for users to post content spontaneously and often anonymously to multiple others, and for those others to then repost that content creates an ideal environment for unverified content to spread rapidly [Nekovee et al. 2007; Derczynski et al. 2015]. This may take the form of rumour or false/malicious information– of which there have been multiple examples in recent times: the incorrect naming of a UK politician in connection with sexual abuse of children [Tweed 2012]; false information about candidates in US political elections [Ratkiewicz et al. 2011]; rumours of volcanic activity following an earthquake in Chile [Mendoza et al. 2010]; and rumours about the location of outbreaks of Ebola [Luckerson 2014]. The increasing use of social media as a news resource by individuals and traditional news media can contribute to the spread of this content [Chei and Long 2012; Gil de Zúñiga et al. 2012].

### 1.1 The consequences of unverified content on social media

The rapid spread of unverified content on social media can cause considerable harm. Rumour or malicious campaigns can be highly detrimental to the reputation of individuals or groups [Tweed 2012]. In times of crisis or unease, social media provides a particularly fertile ground for the spread of rumour [Mendoza et al. 2010], which can then serve to generate further tension amongst affected communities [Burnap et al. 2013; Williams et al. 2013]. In these scenarios, where veracity can be hard to establish, rumours can be repeated and given credibility by traditional media and government agencies such as the emergency services. One key example of this occurred during the civil unrest that took place in English towns and cities in the summer of 2011 [Lewis et al 2011]. An extensive amount of social media activity took place during these events, much of which was picked up and reported on in newspapers, official news websites and television reports. Social media activity was subsequently blamed by the UK government for creating panic and facilitating the spread of civil unrest more widely [Procter et al. 2013a]. As a consequence, Prime Minister David Cameron raised the possibility of blocking networks in order to stop future disturbances [Trenholm 2011] and Her Majesty's Inspectorate of Constabulary [2011a] produced a report on the unrest that called for police monitoring of the 'word on the cyber street' in order to anticipate similar events in future. In another example rumours about the spread of Ebola, noted above, led to panic amongst communities and changes to patterns of behaviour as individuals became fearful of entering spaces where they may been at risk of contracting the virus [Luckerson 2014].

The potential harms caused by the rapid spread of unverified content on social media are described as a global risk factor in a 2013 World Economic Forum (WEF) report "Digital Wildfires in a hyperconnected world" [World Economic Forum 2013]. The report argues that the modern condition of hyperconnectivity provides a platform for massive digital misinformation. This can lead to a 'digital wildfire': the rapid spread of misleading and/or provocative information, with serious consequences. Digital wildfires can threaten the security of individuals, groups and communities, organisations, financial markets, states and even entire populations. The WEF report raises the question of how digital wildfires, and by extension social media platforms, can be governed. In this paper we take up the concept of digital wildfires and aim to set out a research agenda that will help answer this question over governance.

### 1.2 Digital wildfires and the governance of social media

The capacity for rumour and false/malicious information to spread rapidly on social media has inevitably raised questions over whether it is (a) desirable and (b) feasible to enact governance mechanisms to regulate the propagation of these forms of unverified content. This is a highly complex issue which arouses considerable debate centred on several key questions. In the first instance, where does responsibility for governance lie – in the operation of legal codes, through the actions of social media companies or elsewhere? Furthermore what forms of governance might be possible and should they work to deal with unverified content after it has propagated, attempt to slow down the spread of content, or even prevent it being posted in the first place? Some countries, such as China [Blanchard et al. 2013] and Qatar [Gulf Centre for Human Rights 2014], have introduced legal codes that specifically outlaw the online spread of false information and others, including India [Munson 2015], are taking steps to define what can and cannot be posted on social media. Other countries, such as the UK, rely mostly on existing legal codes to regulate social media, meaning that posts are considered in the same way as other forms of communication and may be actionable if they are seen to be defamatory, threatening or indecent [Crown Prosecution Service 2013]. Beyond their deterrent effect, these legal codes are retrospective in nature and deal with social media content after it has been posted – and potentially had harmful consequences. Some governments have taken action to prevent the spread of rumour by blocking access to social media in times of tension – for example, Turkey's use of a temporary court order in March 2015 to block sites including Twitter and You Tube following a hostage situation in Istanbul [Tuysuz 2015]. Such interventions have prompted considerable criticism and raise a final key question: how can the governance of social media be balanced against rights to freedom of speech?

Social media companies themselves are frequently founded on principles of freedom of speech [House of Lords 2014] and the most popular platforms use automated mechanisms to block only the most extreme forms of content. They also require users to sign up to terms of use that set out what kinds of post are acceptable. Posts containing unverified content do not necessarily breach these terms and in any case social media companies have frequently been criticised for appearing to take little interest in regulating content posted on their platforms. For instance Twitter introduced a 'report abuse' button only after vocal and high profile complaints were made about the platform's inability to deal quickly with a malicious campaign against two feminist activists in the UK [Miller 2013; Doshi 2014].

Within these debates over governance, there is a growing emphasis on the need for self-governance amongst social media users. A Select Committee report on social media law in England and Wales [House of Lords 2014] advised users of their responsibilities to monitor what they post and to be aware of how widely their posts can be seen by others. Social media platforms also encourage self-governance in their design. Users often have the option to rate, rank, 'like' or 'favourite' others' posts to indicate that they are worthy of and suitable for reposting etc. They are also able to report malicious posts and/or users to the platform. Additionally, some of the larger social media companies have supported education campaigns to promote responsible posting [WAM 2014; UK Safer Internet Centre, 2015; Pickles 2016].This emphasis on self-governance is echoed in the World Economic Forum report [2013], mentioned above, on digital wildfires. It notes the difficulties of establishing legal codes across countries and the practical complexities inherent to introducing technological mechanisms to govern social media content. It further acknowledges ethical concerns around any restrictions on freedom of speech. Consequently the report points to the value of encouraging social media users to adopt a 'global digital ethos' and behave responsibly when posting. It does not specify what forms this responsible behaviour and digital ethos can, or should, take.

### 1.3 Digital wildfires as a research topic

Given its prevalence, capacity to cause harm and debates over its governance, the spread of unverified content on social media emerges as an important research issue. A key starting point is to consider how we can observe, measure and understand the propagation of rumour and false/malicious information. What disciplines and study methods can contribute to this and what existing research can be drawn on? It is also necessary to look beyond social media content itself and focus on understanding its wider impacts and debates over governance. What are the broader, societal effects of unverified content, and how do individuals, organisations and state agencies verify, contest or dismiss these kinds of content? What are the ethical dimensions of attempts to regulate it? In particular, given the emphasis on self-governance, what self-regulating practices exist and how might they be ethically justified?

To answer these questions, we begin with an overview of current work in this area. We take up the concept of digital wildfires and outline the relevance of a range of work from a variety of disciplines, in particular highlighting a series of insights into the propagation of unverified content. We also show that significant research gaps exist. There is an absence of systematic studies on the effects of social media rumours and false/malicious information, as well as a need to combine empirical research with consideration of how the responsible governance of social media can be determined. We argue that it is necessary to develop a methodological framework that can combine empirical insights on the: (i) spread of unverified content on social media; (ii) the impacts of rumour and false/malicious information; and (iii) the real-time operation of regulatory mechanisms. Furthermore this framework also needs to incorporate discussions and conceptual understandings of the ethical governance of social media. We highlight some recent research that begins to address these gaps and set out an agenda for future research. This agenda establishes a methodology to explore in full the propagation and regulation of unverified content on social media. It promotes high quality research that will inform policy debates over digital wildfires and the governance of digital social spaces.

## 2. THE PROPAGATION OF DIGITAL WILDFIRES ON SOCIAL MEDIA

As described in Section 1, it is well established that unverified content in the form of rumour and false/malicious information can spread very rapidly on social media – creating what can be described as a 'digital wildfire'. Digital wildfires form an important research topic and the starting point for this research is the examination of how this propagation of content occurs. Relevant insights can be found in contemporary work within both computational science and the social sciences.

### 2.1 Insights from computational science

Work conducted within computational science aims to scientifically identify the precise ways that content can spread across social media. Two areas of relevant existing research concern information flow and diffusion. When considering relevant work on the propagation of content, it is helpful to draw a distinction between information flow and information diffusion. Section 2.1 focuses on issues such as the size of the flow (e.g. number of retweets) and its lifetime – attempting to identify models which can be used to estimate such metrics. Section 2.2 then tries to generalise this to a general diffusion process, identifying potential factors that contribute to the overall diffusion process. Clearly both flow and diffusion are related – with the first being a special aspect/facet of the second.

#### *2.1.1 Modelling and Predicting Information flow*

Work within computational science has investigated the predictive factors for the propagation of information flows (i.e. tweets and their retweets) [Lotan et al. 2011]. This can provide insight into how social media content can spread. We define information flow propagation as the process of information spreading to a greater number of people over time. For example, via Twitter through the action of 'retweeting'. Information flows have been measured (i) theoretically, using mathematical models and simulated networks to show that sharing and discovery of information between tightly-coupled nodes leads to the rapid spread of information [Doeer, Fouz and Friedrich 2012]; to investigate 'push' and 'pull' strategies of information exchange in social networks to determine the impact on information diffusion [Chierichetti, Lattanzi and Panconesi 2009], and (ii) using real-world data such as reaction to a terrorist attack, using statistical models to measure the influence of temporal, content and network factors on the likelihood of information propagation [Macskassy and Michelson 2011; Yang and Counts 2010; Burnap et al. 2014b; Burnap and Williams 2015).

Suh et al. [2011] studied retweeting behaviour and identified three factors relating to (i) author profile - number of followers, followees, and tweets; (ii) content features - URLs and hashtags; and (iii) retweets and followers - separating those who have been retweeted a lot and have a large number of followers, from those who tweet a lot and favourite a lot. They built a Generalized Linear Model (GLM) that suggested URL, hashtag and age of account to be most useful for retweet prediction, with follower/followee status being less significant, but still important. Similarly, Zaman et al. [Zaman et al. 2010] used the MatchBox algorithm to predict retweet probability for individual tweets, finding that attributes of the tweeter and the retweeter (similar to author profile of Suh et al.), were most accurate for prediction. Tsur and Rappoport [2012] also investigated Twitter content, specifically hashtags, in the context of the spread of ideas and memes. They identified that the emotive aspects of hashtags were not predictive of the spread of information, perhaps due to their short nature and lack of ‘impact’ in the message in comparison to a longer string of emotive text that has been found to be predictive of information diffusion [Burnap et al. 2014b]. Bandari et al. [2012] undertook the task of predicting the popularity of news stories on Twitter prior to their release. Using classification and regression techniques with features relating to content subjectivity, source and topic, they were able to achieve a reasonable accuracy in predicting a range of propagation likelihood scores, but were less efficient in predicting information flow *size* (number of retweets).

Zaman et al. [2013] used Bayesian models for predicting the number of retweets using a time-series model, predicting at certain points in time as opposed to projecting the final size in the early stages of tweet lifetime. Neither of these models included latent subjectivity and emotion/opinion within the tweet as a feature. Backstrom et al. [2013] identified that temporal factors including the rapidity of comments posted in response to a Facebook status update were predictive of overall thread length. Twitter interaction is slightly different to that of Facebook as retweets are propagated within and between networks of followers and users of shared hashtags etc., as opposed to being visualized in structured “conversations” between “friends” or groups. However, it could be possible to use temporal retweet factors in a similar predictive manner, i.e. to predict total number of retweets using the rapidity of occurrence of initial retweets. Macskassy and Michelson [Macskassy and Michelson 2011] built information propagation behaviour models for Twitter using temporal features such as the time lapse since the original tweet was published and the timing and speed of communication between a tweeter and other users. Both papers suggest that time is an important factor in modelling information propagation and that more investigation is required to use rapidity of retweeting to predict size and survival.

Burnap et al. [2014b] considered the skewed distribution exhibited by retweet behaviour (most tweets do not get retweeted) and used zero-truncated negative binomial (ZTNB) regression method to model retweet likelihood, and Cox regression to estimate proportional hazards to the lifetime of an information flow, for a range of independent measures. They used social, temporal and content factors of the tweet as predictors in both models and found that the sentiment expressed in the tweet is statistically significantly predictive of both size and survival of information flows. The number of offline press reports relating to the event published on the day the tweet was posted was a significant predictor of size; the tension expressed in a tweet was also a predictor in relation to survival, i.e. the duration of the information flow as measured by the time between the first and last retweet. These findings suggest that the media, social tension, and information flows are interrelated in some way. Furthermore, they found time lags between retweets and the co-occurrence of URLs and hashtags also emerged as significant, suggesting that (i) the quicker people engage with a tweet, the more likely it is to ‘go viral’, and (ii) aides to discoverability and links to additional information (e.g. evidence) are an important predictor of information flow. On the particular topic of self-governance, Burnap and Williams [2015] found that *cyberhate* – hateful or antagonist tweets targeted at social groups based on their personal attributes – were statistically less likely to propagate and become a large information flow following an event where wide scale public distribution of such content could pose a risk to public safety (a terrorist attack in Woolwich, UK). This study provided some of the first evidence to suggest that users of social media sites, Twitter in particular, do self-govern by refraining from propagating socially unacceptable material.

### *2.1.2 Factors influencing Information diffusion*

In general, information diffusion and spread in social networks (as exemplified through social media systems, e.g. Facebook, Flickr, Twitter, Instagram, etc.) can be divided into four components: actors (i.e. participants involved in the exchange), content (what is contained in the exchanged messages), underlying network structure (connectivity between participants or derived relationships across a number of metrics related to content or actors), and the diffusion process itself (based on some of the factors identified in section 2.1.1). Given our interest in understanding the issues related to governance in 'real world' systems, we focus on the existing literature that has reported real-world empirical research on information diffusion, rather than simulated or theoretical research. Current research in this area suggests that content of posts, author profile and temporal issues are all relevant to whether or not a post is propagated. Each of these components – along with their contribution towards the study of the propagation of unverified content – are briefly described below:

*Actors*: A number of different models have been developed that describe how a certain fraction of users decide to follow a particular course of action (based on the behaviour of their "friends" on-line). Such "threshold" models (influenced by a threshold number of "friends" who follow an action to influence behaviour/adoption by an actor) [Watts and Dodds 2009] describe the perceived benefits seen from the perspective of an actor based on the influence exerted by friends in a social network [Morris 2000], taking account of on-line relationships between the actor and friends. Various studies have demonstrated the existence of such a "threshold" in social and behavioural contagions online [Romero et al. 2011], i.e. the probability that an additional positive signal will trigger adoption depends extremely sensitively on how many other signals have been observed (regardless of the order in which they were observed): just below the threshold, a single observation can increase the adoption probability from near zero to near one, where otherwise it will have little effect. Each individual within the population may have a different threshold (capturing the relevant psychological attributes of this individual with respect to the particular decision at hand). The principle of homophily (measuring similarity between users) also has significant influence between the likely actions of users on-line, as demonstrated through various research undertaken in the development of recommender systems. The homophily effect was suggested to greatly promote behavioural contagion other than the peer influence [Aral et al. 2009].

*Content*: Significant research has focused on the *innate appeal* of content to influence users to share it. A number of features can be extracted from Twitter content, such as hashtags, number of words, spelling, lexical items, location in tweets, in addition to (often subjective) emotional and cognitive aspects, to predict likelihood of retweeting the message. Additional aspects include detection of topic and locality. In the context of social media sites like Twitter, topic locality refers to the assumption that semantically similar hashtags are more likely to be mentioned in the same posts and therefore to be close to each other in the hashtag co-occurrence network. Associating topic locality with user interests can also be the basis to influence retweets, i.e. determining whether a user would be interested in a newly arrived message is estimated by the similarity between the user interests and the message. [Java et al. 2007] looked into communities of users in the reciprocal Twitter follower network and summarized user intent into several categories (daily chatter, conversations, information sharing, and news updates); a user could talk about various topics with friends in different communities.

*Network Structure*: Network structure captures the underlying topology of linkages between individuals, which can influence how quickly a message will propagate [e.g. Albert and Barabasi 2002; Barrat et al. 2008]. A variety of network structures exist that may be used as a basis for capturing such a topology, e.g. a random network, scale-free network, small world network, etc. Social networks naturally consist of communities corresponding to certain social circles or interest groups, differentiating social networks from other kinds of network structures - such as biological (e.g. metabolic and protein interaction, etc.) networks and technological networks (e.g. train routes, telecommunications interconnectivity, energy/power grids, etc. [Newman and Park 2003]. Although there is no clear consensus, a community is often defined as a densely connected subgraph. Community detection in arbitrary network structures often provides a useful basis for determining the likelihood of message propagation. The rise in the use of social networking sites has motivated the study of network structure on rumour detection [e.g. Kwon et al. 2013] and propagation [e.g. Dechun and Chen 2011]. Several theoretical studies of the influence of network structure [e.g. Kostka et al. 2008; Chierichetti et al. 2009; Doerr et al. 2012] have demonstrated that rumours spread more quickly in social networks compared to other classical network typologies. Related work has focused on theoretical studies of rumour control strategies for social networks [e.g. Bao et al. 2013].

*Diffusion Process*: this models how information gets transmitted amongst a group of users. A Twitter message, for instance, can pass from one individual to another through social connections and "infected" individuals (many such processes are based on the "epidemic model" – which models the spread of disease within a population) can, in turn, propagate the information to others, possibly generating a full-scale contagion. Early diffusion models used in social networks were strongly influenced by epidemic models. They were later extended to include cascade phenomena [Goldenberg et al. 2001], factors that influence the speed of spreading such as information recency, patterns of connectivity and message exchange between individuals, the existence of clusters of users (based on homophily), etc. Recent work on the analysis and modelling of online information diffusion aimed to reproduce statistical features of the cascades as in the empirical data or learn the mechanism of how a message is propagated.

Guille and Hacid [2012] developed a model to predict the diffusion of information in online social networks, focusing on social, temporal and content factors. They identified a Bayesian logistic regression as a favoured predictive model. While the model performed well for diffusion, it was less effective at predicting size, indicating that predictive features of information diffusion and information flow size are independent

Yang and Counts [Yang and Counts 2010] constructed a topic-based diffusion model based on user mentions, where a mention constitutes the propagation of information from one user to another. They aimed to predict speed (time taken to reach first mention), scale (the number of first-order mentions of the user), and reach (the number of hops the mention produced). They employed the Cox proportional hazards regression model to quantify the degree to which features of the tweet or the original tweeting user were useful in predicting speed, scale and reach. They found that predictive features varied across topics for speed and scale with the amount a user was mentioned in the past emerging as most predictive.

Work within computational science helps to establish how and why the kinds of content associated with digital wildfires propagates on social media. The research on information diffusion outlined above can also form an important bridge to social scientific work in this area. This is because it identifies actions and measurable concepts that include a social element and that can help to explain the spread of content both quantitatively and qualitatively. This is discussed next.

### 2.2 Insights from the social sciences

Social media is emerging as an important topic within the social sciences [Trottier 2012; Murthy 2012a; 2012b; Lupton 2015]. In particular, attempts have been made to reach a theoretical understanding of the position and role of social media in modern societies. The rapid and widespread uptake of platforms such as Twitter, Facebook etc. is characterised as a significant innovation that has created new ways for people to interact and to share information. These online spaces can be conceptualised as a socio-technical assemblage that creates a new public sphere [Mossberger et al. 2008]. Social media streams can therefore be considered as new sources of information on the perceptions, opinions, actions, feelings and tensions expressed by individuals and their communities. They also have the capacity to transform social relations and to shape, and be shaped by, governments and other institutions. In addition to these theoretical understandings, social scientific study methods offer a means to trace the inter-relationships between social media content and other phenomena. As an example, McEnery et al. [2015] investigated the content of social media and traditional media items in response to an ideologically motivated murder in the UK. They analysed corpora of mainstream press coverage and Twitter coverage of the event in order to identify influences on and between the two forms of media. They found that mainstream media sources were a strong presence on Twitter, with users frequently posting links to sites such as bbc.co.uk etc. Twitter posts played a key role in the initial reporting of the murder but as the event unfolded mainstream media became less likely to pick up on content posted on the platform. Despite the observable influence mainstream media exerted on social media, users frequently presented interpretations of information drawn from newspapers etc. rather than straightforwardly repeating it.

Social scientific approaches can contribute to understandings of digital wildfires in two key ways. Firstly, the social sciences have a long history of investigating rumour as a form of collective behaviour [Allport and Lepkin 1954; Shibutani 1966; Goffman 1981; Dingwall 2001]. This body of work identifies the societal functions played by rumour. For instance rumours can serve to fill information gaps and knowledge deficits, and establish shared narratives as a kind of sense-making device – in particular in times of great uncertainty. Further contributions can be found in work on moral panics [Cohen 1973], the amplification of deviance [Wilkins 1967] and the self-fulfilling prophetic qualities of rumour [Innes 2004]. These insights can be applied to the analysis of social media. They can identify ways to understand online rumours in terms of the purposes they serve, the kinds of 'signal events' (such as crime and civil unrest) that might cause rumours to spark and the social conditions in which they might propagate. Crucially, this approach offers a way to move beyond the notion of 'memes' [Dawkins 1989 p. 368] – in which rumours can be seen as units of cultural transmission, replicating as they spread from post to post – and towards generating a better understanding of the relationship between wider social structures and people's conduct online.

Secondly, the social sciences (alongside other disciplines such as linguistics, cultural studies and pragmatics) also have an established history of conducting in-depth analyses of the content and conduct of communicative acts. Arising from a variety of analytic traditions this work highlights the ways in which communication is a social phenomenon. For example, discourse analytic studies [Coulthard 1977; van Dijk 1985] describe spoken and written communication as drawing on social understandings that can often serve to reinforce established cultural meanings and power relations. Work conducted within interactionist approaches [Sacks et al. 1974] describes the ways that communicative acts develop sequentially by responding to what has come before and projecting what kind of response might follow. Once again, relatively little work has been done so far that focuses specifically on the communicative acts associated with the propagation of digital wildfires. A notable exception is Procter et al.'s study of rumour propagation during the riots in England of 2011 [Procter et al. 2013a]. This examined how rumours spread on Twitter, focusing in particular on how individuals made sense of information of uncertain quality and provenance. This pioneering work involved analysis of the event and context-specific qualities of speech acts, including the presence and powers of counter-speech. It is described further in the next section. More broadly, a relatively substantial amount of existing work has analysed phenomena such as online 'trolling' (i.e. knowingly making inflammatory comments) and 'hate speech' (content likely to incite violence or prejudicial actions against individuals or groups) as social activities. Hardaker [2010] argues that the anonymity afforded by social media platforms reduces the extent to which users can be held accountable and open to censure for their posts. This creates a fertile environment for a variety of communicative acts such as impoliteness, displays of aggression and disruption of interaction – all of which can be characterised as trolling behaviours. In a study of inflammatory comments made on You Tube, McCosker [2014] describes the comments field as a participatory space drawn on by users to express their identification with a specific place, nationality or culture etc. In this context disagreements are treated as provocative and interactions between users can become increasingly passionate and vitriolic as expressions of national, social or cultural citizenship are made and defended. In two final examples, Awan [2014] and Williams and Burnap [2015] analysed anti-Muslim hate speech on social media. Awan created a typology to characterise the different kinds of content being posted and suggest user profiles of those posting the content. Williams and Burnap [2015] built a classifier for anti-Muslim hate speech and modelled its propagation following the Woolwich terror attack, finding far right groups were most likely to spread hate within the first 36 hours following the event.

As social media is a relatively new area of focus for the social sciences, a number of significant research gaps exist. The social sciences face the challenge of establishing a framework to systematically investigate and understand the impacts of social media on modern society and the impacts of modern society on social media. It is particularly important to examine the inter-relationships between social media, social media content and societal dynamics and structures. Social scientific study methods position social media within a broader societal context; they also produce a sophisticated understanding of the functions played by content such as rumour and recognise the different roles and interests of individuals and groups involved in the posting and spread of content. They provide conceptual and methodological tools to examine in detail the content of social media posts, identifying the communicative acts performed within individual posts and analysing the relationships between posts as content propagates. We argue that, to date, social media use has been treated as if the phenomenon and how it is used is already understood [Tolmie et al. 2015]. Here, Conversation Analysis [Sacks et al. 1974] has much to offer in providing a basis for understanding how interaction is actually accomplished through social media and thus providing the foundations for investigating more complex social phenomena that depend upon it.

The challenge for social sciences is to find ways of harnessing computational analytics so that its study methods are commensurate with the large and ever growing volumes of social media data.

### 2.3 Combing approaches and ways forwards

This overview has set out to demonstrate the ways that state of the art approaches in computational science and the social sciences may contribute towards understanding the propagation of unverified content on social media. Work within computational sciences on information flows and diffusion points to specific factors for propagation that relate both to the content of a post and to features such as network structure, timing of posts etc. Work on information flow motivates further work that looks more closely at the content of posts. Work on diffusion is likely to be highly relevant to understanding how rumour and false/malicious content spreads rapidly over social networks. The application of insights from the social sciences can produce a sophisticated understanding of the functions played by digital wildfires, the societal contexts in which they occur and the social structures that underpin them. Social scientific study methods can also provide a detailed characterisation of social media content, the actions it performs and stimulates, and the ways in which people engage with it [Tolmie et al. 2015]. Conversation Analysis, for example, points to the importance of looking at 'conversational threads' as a way of enriching the notion of information flow in ways that transcend the individual tweet and enable identification and modelling of its interactional and dialogical features, which is the approach being pursued by the Pheme project[1] [Zubiaga et al. 2016]. The challenge is how to create an inter-disciplinary fusion, so that social science insights may contribute to computational analysis and computational science tools may make the analysis of large amounts of social data tractable [Ruppert et al. 2013].

The above discussion of existing research points to particular ways in which computational analysis can be combined with social scientific insights to respond to the challenges of digital wildfires, their consequences and opportunities for governance. Of particular importance is the systematic qualitative inspection of social media data, such as Twitter, to generate inductively typologies of action and agency, including the identification of specific dialogical features. This creates small-scale annotated datasets, which can then be used to train machine learning algorithms to recognise content of interest in much larger datasets [Housley et al. 2014; Zubiaga et al. 2015b]. More broadly, we would argue for an integrated 'research workflow' that combines computational science and qualitative social science methods to examine how

[1] www.pheme.eu

people use of social media *in situ* [Tolmie et al. 2015]. This approach, we argue, is essential to inform governance, regulation and intervention practices. We set out a research agenda for this integrated research workflow in Section 5 of this paper.

## 3. RESPONSES TO DIGITAL WILDFIRES BY INDIVIDUALS AND AGENCIES

As the use of social media has grown exponentially, so too have the efforts of certain agencies to respond to and manage this kind of content. The capacity for digital wildfires to cause significant harm has inevitably led to attempts control this kind of content and debates over the appropriate regulation of digital social spaces. Therefore in addition to understanding why and how digital wildfires propagate, it is also necessary to consider how individuals, agencies and communities respond when they appear on platforms such as Twitter. Work in this area deepens understanding of the potential impacts of digital wildfires and consequences for governance.

A range of different agencies have an interest in responding to unverified content posted on social media. Prime examples of such groups are governments and law enforcement agencies. As discussed in Section 1.2 a number of countries have taken steps to criminalise the posting of unverified content or even block access to social media at certain times. In 2015 a number of social media users in China were imprisoned for making posts about the country's stock market downturn that were described as 'false' and 'destabilising' to the market. [BBC News 2015]. In the UK a 2011 report published by Her Majesty's Inspectorate of Constabulary [2011b] implicated social media sites in the fomentation of unrest and public disorder – citing, for example, protests about public spending cuts and tuition fees in education. As a consequence, police interest in social media is increasing. In addition to pursuing potential illegal acts committed through social media, policing agencies in the UK are becoming more interested in the use of social media to gather intelligence [Procter et al. 2013b]. This intelligence gathered may focus on signs of tension within or between certain communities and the spread of rumour about individuals or events and once gathered can be drawn on to assist in safeguarding the public. Social media platforms also provide the police and other authorities with the capacity to dispel rumour and 'reassure' the public in an attempt to engineer social order.

Other agencies with a stake in responding to digital wildfires include the emergency services and mainstream media outlets. The ability for social media content to propagate rapidly means that crowd-sourced 'citizen journalism' reports on critical events etc. can quickly gain a high profile and become influential well before other agencies such as mainstream media, emergency services and the police etc. have been able to mobilise and react. Consequently this unverified content can be picked up by these agencies and become part of their own response. Where this content is incorrect, this can have very negative effects. Wendling et al. [2013] describe crowd sourcing efforts following the earthquake in Haiti in 2011. The aggregation of social media data concerning building damage etc. facilitated development of an interactive map that was used by search and rescue teams to find survivors. However, the overall accuracy of the aggregated data was weak and ultimately its use led to a misdirection of aid resources. Another example is the 'amateur sleuthing' on platforms such as Twitter and Reddit following the Boston bombing of 2013 [Davison 2013]. The (false) identification of a missing university student as one of the bombers was picked up by mainstream media and led to the focusing of attention on his family. As these cases indicate, the unquestioning acceptance of unverified content by media and government agencies brings a number of kinds of societal risk. This is now being countered by efforts from mainstream media, emergency services, volunteer crisis responders etc. to track and verify content appearing on social media and create their own social media presence to facilitate and encourage more accurate posting [Derczynski et al. 2015; Procter et al. 2013b; Wendling et al. 2013].

A final response type concerns social media users themselves. User self-governance has been highlighted as central to the appropriate regulation of social media and can be observed in a number of forms. Users might respond directly to counter malicious posts containing hate speech or 'trolling' [Gagliardone et al. 2015] or forward the post on to a sentinel site such as 'Yes, you're racist' or 'Yes, you're homophobic' etc. which set out to draw attention to (as their names suggest) specific kinds of hate speech. These sites make the original post available to a wider audience with the intention of shaming the poster or even – as in the case of the site 'Racists getting fired' – encouraging others to identify the poster in real life and affect negative consequences on him/her, such as the loss of employment. Relevant research has been conducted on the ways that users respond to rumour online. This work has already been referred to in this paper and concerns the analysis by Procter et al. [2013a] of social media activity during the England riots of 2011.

The 2011 riots began as an isolated incident in Tottenham, London, on 6 August but subsequently quickly spread across London and to other cities in England [Lewis et al. 2011]. In their immediate aftermath, some politicians and media commentators were quick to blame social media for the scale and extent of the disorder. It was claimed, for example, that social media were used amongst rioters to organise their activities as well as to incite others into unlawful acts, inflame tension and spread panic through the posting of unsubstantiated rumours. This lead to calls for social media to be closed down during such events. Procter et al. [2013a] concluded that the evidence pointed overwhelmingly to Twitter being used for more positive ends during the time that disorder was spreading (for example, in the form of anti-riot and pro-police messages) and, in particular, in the mobilisation of volunteers to clean up after the riots. This view has been supported by a number of other studies [Casilli and Tubaro 2011; Baker 2012; Bassell 2012; Tonkin et al. 2012]. The differences between immediate political claims about the influence of social media posts during the riots and subsequent research findings demonstrates that the understanding of the impacts of social media content can be a highly contested area. It further demonstrates the value of scrutinising these kind of claims through careful analysis, with the inclusion of close attention to the content of posts.

Procter et al. [2013a; 2013b] examined in detail a number of rumours that were propagated through Twitter during the riots. One was a rumour that a mob of rioters was attacking Birmingham Children's Hospital. This example reflects a pattern or trajectory common to the different rumours studied.

1. A rumour starts with someone tweeting about the occurrence of an alleged incident.

2. The rumour is picked up by their followers and gets retweeted. Some form of evidence – eyewitness reports, references to mainstream news sources, links to pictures or to mainstream news sources on the Web, etc. – may be added as the original tweet gets retweeted and various reformulations of the rumour also begin to appear.

3. Others begin to challenge its credibility (i.e. make a counter-claim), assessing the evidence and offering refutations of it, perhaps on the basis of logical arguments (e.g. "it's not possible because …") or new information that throws into doubt the veracity of evidence previously offered.

4. A consensus begins to emerge. Where this is that the rumour is false, it may nevertheless re-surface in the corpus as latecomers pick up the original tweet and join in.

The use of links to other media, e.g., mobile phone images, blogs and online newspaper sites as corroborating evidence is another common feature in all of the rumour case studies conducted by Procter et al. However, they show that this evidence cannot always be taken at face value. For example, the authenticity of the image purporting to show that the London Eye burning was subsequently challenged by claims that it had been faked ('photoshopped') to give the impression of a blaze [Procter et al. 2013a]. This and subsequent cases, such as images circulating across social media during Hurricane Sandy [Gupta et al. 2103], make it evident that numerous individuals are prepared to go to considerable efforts to convince others of the credibility of rumours they know to be false. These same case studies also indicate that social media users can – and do – draw on a range of resources to challenge and discredit false rumours, revealing the key role of natural correction mechanisms for rumour debunking. The Pheme project is investigating how computational methods may be used to amplify these mechanisms [Derczynski et al. 2015; Zubiaga et al. 2015a]. It is also possible to observe instances of user self-governance in response to malicious social media content. Again, an important question for future research is how such self-governance processes, including the efforts of sentinel sites, might be amplified.

### 3.1 Remaining questions

As described above, a range of agencies have an interest in monitoring, responding to and shaping the spread of unverified content on social media. Individual users can also play a role in this response, for instance by countering malicious posts or challenging the credibility of rumours. The success of individuals and agencies in verifying, challenging or limiting unverified content can vary.

Understanding the responses of individuals and agencies needs to be considered an important component of research into digital wildfires. The issues discussed above raise further relevant research questions. For example, what other kinds of agencies take an interest in digital wildfires and what steps to do they take to respond to them? How do the actions of different agencies interweave in a given scenario – such as in response to civil unrest or a natural disaster? Furthermore what tensions might emerge between traditionally authoritative sources of response (mainstream media, emergency services etc.) and the newly emergent 'citizen journalism' on social media? Collaborative efforts by large numbers of 'produsers' [Bruns 2006] can provide coverage of events that competes with mainstream media and attempts to set an agenda for the police, emergency services and other agencies. What are the consequences of these competing forms of coverage – for example, in terms of public trust and the management of response to critical events?

The effective and appropriate governance of social media emerges as an important issue. How can – or should – digital wildfires be regulated? As discussed in Section 1.2 some legal codes do cover social media posts, as do the Terms of Use of individual platforms. In addition, the efforts made by mainstream media, law enforcement agencies and the emergency services etc. to track, verify and influence social media content displays an assumed governance role – not only in responding to content but also in influencing it. User behaviours can also play a role in governance. These various groups have an interest in managing digital wildfires but to what extent is this appropriate? Actions to challenge, limit or halt digital wildfires are based on the assumption that when unchallenged they can cause harm. But how does that harm balance against the positive roles that can be played by social media and would any attempts at governance limit those beneficial impacts as well as the negative ones? Is it possible that governance mechanisms could also cause harm – for instance by excessively punishing individuals or limiting freedom of expression? It is crucial that investigation of digital wildfires on social media includes consideration of the ethical dimensions of governance.

## 4. ETHICAL DIMENSIONS OF DIGITAL WILDFIRES AND SOCIAL MEDIA GOVERNANCE

As we have seen, certain responses from different agencies and individuals to the spread of rumour or false/malicious information constitute forms of governance that attempt to shape in some way the spread of this content on social media and/or its (assumed) harmful impacts. The repeated occurrence of social media rumours and malicious campaigns etc. could be taken to suggest that the effectiveness of these existing mechanisms is limited and that further governance is needed to protect individuals and populations from the harms of digital wildfires. This could take a number of forms, for example: stricter/new legal codes, changes to the Terms of Use on social media platforms; technical mechanisms to limit the speed of content propagation in some scenarios; the provision of 'lie' buttons or esteem factors on social media to enable users to indicate whether posts are credible or creditworthy; or education campaigns to encourage responsible posting amongst users.

Attempts to prevent or manage digital wildfires through further regulation are founded on the assumptions that the spread of unverified content can be harmful and that these harms need to be reduced. However, where governance attempts to limit the spread of content, this inevitably also limits users' ability to post and access information. This creates tensions around censorship and freedom of speech. Attempts to limit the harms caused by digital wildfires might risk also limiting certain benefits provided by social media or risk causing harm themselves. These issues are recognised in the World Economic Forum [2013] report on digital wildfires:

> "Establishing reasonable limits to legal freedoms of online speech is difficult because social media is a recent phenomenon, and digital social norms are not yet well established. The question raises thorny issues of the extent to which it would be possible to impose limits on the ability to maintain online anonymity, without

> seriously compromising the usefulness of the Internet as a tool for whistle-blowers and political dissidents in repressive regimes."

As this quotation highlights, consideration of governance mechanisms in relation to digital wildfires raises a complex ethical landscape. In addition to robust understanding of how existing mechanisms play out in practice and what gaps in regulation exist, informed decision making about social media governance requires close attention to ethical concepts and debates. Here we give a brief overview of two fields that can offer valuable insights in this area.

### 4.1 Computer Ethics and Justification for Governance

The inter-disciplinary field of computer ethics has developed in response to the realisation that computing technologies have the potential to affect our lives in ways that interfere with our preferences, duties or responsibilities [Johnson 1985; Adam 2001; Bynum and Rogerson 2003; Floridi 2010]. Theoretical positions within computer ethics include traditional ethical theories from moral philosophy – namely, deontology, teleology and virtue [Stahl 2012]. These positions provide a means to consider whether social media governance mechanisms to manage digital wildfires can be ethically justified. Questions over harm and truth are central. Governance measures to reduce the harm caused by the spread of unverified content are based on a teleological positon, which focuses on the consequences of actions. Where the aggregate measure of disutility of an action – such as the rapid spread of an online rumour – outweighs its aggregate measure of utility, this can provide a justification for governance. In the case of the governance of digital wildfires, important questions arise over how the harmful consequences of rumour etc. can be measured and assessed. Is harm an objective entity or does it include subjective components such as the feelings of people affected? Over what timescale should consequences be measured, and is it possible that short term harm can lead to longer term benefit? Do the harms caused by digital wildfires outweigh the harms caused by limiting freedom of speech?

In the teleological position truth is of secondary importance to consequence. So the truthfulness or otherwise of an online post matters only in relation to the consequences it has. By contrast the actor's intention to be truthful is central to the deontological position. This perspective is based on the assessment of the actor's intent, meaning that an intention to speak the truth can provide ethical justification for the online posting of unverified content. The virtue perspective takes a similar position, in that the speaking the truth is a virtue and spreading untruths or malicious content is a vice. These perspectives problematise issues of governance because it becomes necessary to attend to the truthfulness or otherwise of unverified content plus whether or not the individuals involved in starting or promoting the content believe those claims to be true. Of course determining the intent of users in posting content is extremely difficult. Furthermore it is highly unlikely that the multiple users involved in the mass spread of online unverified content share the same intentions.

### 4.2 Responsible Research and Innovation

Computer ethics raises important conceptual questions over governance but provides little in the way of practical guidance on the forms that different governance mechanisms could/should take and how they can be applied. A means to overcome this absence arises in work produced within the field of Responsible Research and

Innovation (RRI) [Owen et al. 2012; Stahl et al. 2014]. RRI has gained prominence in recent years as an EU initiative but academic work within this field also includes approaches developed in the US and worldwide. This work typically focuses on scientific research rather than ICT but the attention given to practices of 'responsible development' and responsible governance' [Kjølberg 2010; Roco et al. 2011] have relevance to consideration of social media. For instance, these approaches are sensitive to the huge challenges represented by timing governance and responsible behaviour to attend to quickly changing environments. This can provide insight into governance relating to the rapid spread of content on social media. Furthermore these approaches are also sensitive to local, social and cultural contexts. These contexts matter given the different kinds of social media, their varying outreach and their entanglement with offline actions and events. The RRI approach offers insights into the practical application of ethically justified social media governance. In particular, an RRI approach suggests that heedful forms of governance may need to account for existing forms of self-governance amongst social media users, be reflexive of their own possible shortcomings, be continually scrutinised in the light of new digital challenges and be realistic in their ambitions. Therefore in addition to formal foresight and prediction methods it may be necessary to adopt anticipatory governance measures that involve different stakeholders and allow for the co-construction of 'desirable futures' [Stilgoe et al. 2013: 1571).

## 5. ADDRESSING GAPS IN THE RESPONSIBLE GOVERNANCE OF SOCIAL MEDIA: A RESEARCH AGENDA

So far in this paper we have demonstrated that the propagation and regulation of digital wildfires form important topics for research and have conducted a review of existing work in this area. We have described the ways that they can spread rapidly across social media and take the form of digital wildfires that have significant negative consequences for individuals, groups, organisations and communities. We have also outlined the debates over the governance of social media that have arisen in recent years. Our literature review identified work from both computational science and the social sciences that can contribute to understanding the propagation of unverified content on social media and also pointed to the analytic benefits of combining approaches. We described the potential for social scientific work to illuminate the broader societal contexts of digital wildfires and conduct detailed examination of the content of social media posts. We discussed existing research on the responses of individuals and different agencies to unverified content and highlighted its relevance, alongside ethical considerations, to questions over the appropriate regulation of social media. Our discussion of this literature has led to two key findings: (1) the benefits of an interdisciplinary approach towards the examination of social media content and its propagation; and (2) the value of connecting analyses of social media content with examination of the responses of different individuals and agencies and questions over responsible governance.

Despite the relevance of this existing work, there is a need for further research in this area. We have identified a number of significant research gaps that need to be filled. These concern: (1) qualitative analysis of social media posts containing rumour and false/malicious content to inform computational analysis; (2) social scientific analysis of the inter-relationships between social media and society, in particular the impact of digital wildfires on individuals, groups, organisations and communities; (3) examination of the ways that various responses and governance mechanisms related to digital wildfires play out and interconnect in different scenarios; 4) combining empirical findings with conceptual understandings to produce a framework through which the responsible governance of social media can be addressed.

In this section we put forward a research agenda that takes up the challenge of combining the empirical examination of digital wildfires with exploration of responsible governance. This agenda has the capacity to begin to fill existing research gaps and inform policy debates.

### 5.1 A research agenda

In order to take up the research challenge presented by the contemporary phenomenon of digital wildfires, it is necessary to *build an empirically grounded methodology for the study and advancement of the responsible governance of social media.* This requires an *interdisciplinary* approach that incorporates relevant contemporary developments in computational science, the social sciences, computer ethics and responsible research and innovation.

Empirical work is necessary to describe and analyse *the communicative affordances* of social media in terms of how these platforms facilitate the posting and *real time propagation of content* associated with digital wildfires. This involves examining the temporal structures of information flow and diffusion on social media. As described in Section 2.3 combining study methods from computational science and the social sciences offers particular benefits here. The systematic qualitative inspection of social media data can generate typologies of action and agency within digital interaction relevant to digital wildfires. These may include forms of claims making, legitimation, rebuttal, information sharing or agreement etc. which can then be applied to larger datasets using computational approaches. Qualitative analysis involving the systematic manual inspection of social media streams and content can capture the relational and interactional dynamics of Twitter and other social media platforms. Drawing on interactional approaches within the social sciences these enable empirically grounded interaction feature identification which can then inform the generation of coding frames for the crowdsourcing of annotation and coding for machine learning. Given the significance attributed to user self-governance during digital wildfires, it is further necessary for this analysis of the communicative affordances of social media to focus on practices of self-regulation. Drawing on the inter-disciplinary approach outlined above it is possible to identify practices of self-regulation and determine their impact on the propagation of content.

Moving beyond social media content itself, further empirical work is required to *recognise the relevance of different stakeholders* in digital wildfire scenarios. As discussed in Section 3 there are a range of agencies – law enforcement, traditional news media etc. – with an interest in verifying, contesting or dismissing rumour or false/malicious information as it appears on social media. It is necessary to examine how these agencies interpret the communicative affordances of social media and responding to them accordingly in real time scenarios. The use of ethnographic approaches (based on observational fieldwork and interviews as outlined in Section 2.3) provides an ideal means to capture and examine these practices in order to aid understanding of how different agencies manage and respond to the threats generated by social media in real time. In combination with findings from the analysis of social media content, ethnographic work also provides an empirical foundation to support conceptual social scientific work on the broader societal contexts of social media. This makes a significant contribution to understanding of *the relationships between online content and other behaviours, actions and events*. For example, it can help to identify the different kinds of harm – subjective, measurable etc. – caused by digital wildfires to individuals and groups and support work to conceptualise the consequences of these harms for social media governance.

Finally, it is necessary to *explore fully the different opportunities for the regulation of social media in relation to digital wildfires*. As described above, empirical analysis of social media content can identify practices of user self-regulation and their effectiveness in specific digital wildfire scenarios. It can also determine the limitations of self-governance and identify instances in which intervention by other agencies may be practical and beneficial. In addition, full exploration of governance requires seeking out the different viewpoints of relevant stakeholders on the effectiveness of current mechanisms and opportunities for further forms of governance. It is also necessary to acknowledge the ethical dilemmas and controversies that arise over social media regulation and to seek to understand different viewpoints concerning the ways that digital wildfires can or should be prevented, managed or limited. As described in Section 4.2, the adoption of insights from responsible research and innovation can inform practical advances in the ethically justified governance of social media – with particular regard to maximising the potential for user self-governance. They can help to incorporate evidence of digital wildfires and their consequences with broader understandings of the societal contexts in which digital wildfires occur and different views on governance. This supports the development of governance structures that are effective and ethically sound whilst acknowledging the benefits of social media and the potential for future challenges to occur.

This research agenda provides the foundation for robust analysis that advances knowledge of social media, social media content and social media governance in relation to digital wildfires. It will contribute to understandings of opportunities for ethical governance via the actions of individuals and different agencies. This agenda marks a way to fill significant existing research gaps and can benefit the research community by highlight new methodologies and tools for harnessing the potential of social media data. It also informs policy debates over social media governance and can benefit a range of key groups. These include: policy makers with formal responsibility for developing digital society initiatives; government agencies responsible for social media policy implementation and governance processes; voluntary sector organisations and other groups involved in the promotion of responsible social media behaviours and of social cohesion etc.; and vulnerable social media users and their advocates (for example, school students and their teachers) with an interest in protecting them and advancing their digital maturity.

### 5.2 The “Digital Wildfire: (mis)information flows, propagation and responsible governance” project.

The authors of this paper are currently working on a research project that pursues the research agenda outlined above. The “Digital Wildfire: (mis)information flows, propagation and responsible governance” project is an inter-disciplinary study that seeks to advance empirical knowledge of the propagation of unverified content in the context of digital wildfires and inform debates over the effective and ethically justified governance of social media. We are undertaking a variety of research activities to achieve the requirements set out in the research agenda. These involve:

(1) *Scoping ethical questions in relation to digital wildfires*; Drawing on major traditions and concepts in computer ethics [Stahl 2012] we identify the core ethical issues arising from the phenomenon of digital wildfires. For example, the truthfulness (or otherwise) of social media posts and the intent of users when propagating content. We then reflect on the strength and limitations of these competing traditions of ethics in relation to the governance of digital wildfires.

(2) *Scoping existing governance mechanisms, their limitations and possibilities for further mechanisms*; We identify key existing mechanisms relevant to the governance of digital wildfires; namely, the law, social media platform governance, institutional regulation, and user self-governance. We review the capacity for each mechanism to manage, limit or prevent the spread of content in a digital wildfire scenario and identify possibilities for further governance mechanisms [Webb et al. 2015].

(3) *Conducting case studies of digital wildfires through the quantitative and qualitative examination of social media datasets*. We apply the research approach advocated in Sections 2.3 and 5.1 by combining qualitative and quantitative techniques to examine social media datasets. We analyse information flows during digital wildfires, with particular attention to the occurrence of self-governance practices such as counter speech and their implications for the spread of content.

(4) *A Delphi panel* [Adler and Ziglio 1996]. We conduct a series of questionnaires to gather stakeholder opinion on ethical and governance issues relevant to social media. Participants come from 4 groups that reflect the different governance mechanisms identified in activity 2: law, social media platforms; institutions and social media users. They submit answers to a series of open-ended questions which seek their opinion on the appropriate regulation of digital social media and digital wildfires. They then have an opportunity to respond to each other as further rounds of questionnaire seek to identify areas of consensus.

(5) *Ethnographic interviews and observations*. We undertake fieldwork at different sites where agencies and organisations have an interest in dealing with (potentially) negative consequences of the spread of social media content containing rumour or false/malicious information. For instance, police control rooms, anti-harassment organisations, law enforcement agencies, education agencies etc. We gain understanding of the procedures and activities undertaken by these groups and the challenges they face when dealing with tensions, conflicts or disturbances that might arise from this content.

(6) *Ethical security map*. We draw together our study findings to produce an ethical security map for social media stakeholders. This will take the form of a practical tool to help different users navigate through social media policy and aid decision making with regard to the spread of content.

Other project outputs include the development of a training module on digital maturity and resilience for use in secondary schools and the production of artwork to promote a creative understanding of digital wildfires amongst a broad range of audiences.

## 6. CONCLUSION

The contemporary popularity of social media platforms creates a condition of hyperconnectivity in which users can share content with multiple others spontaneously. This enables the swift spread of content and risks ‘digital wildfires’ in which rumour or false/malicious information propagates rapidly and causes considerable harms. Governance debates emerge over how digital wildfires can be managed, limited or prevented. High quality research can advance knowledge regarding unverified content on social media and inform debates over the effective and responsible regulation of digital social spaces. This research requires an interdisciplinary approach that combines the empirical analysis of social media content and propagation with the examination of responses of different individuals and agencies to digital wildfires as well as attention to questions over responsible governance. It is necessary to overcome gaps in existing research in order: to analyse the communicative affordances of social media; recognise the relevance of different stakeholder perspectives and experiences in digital wildfire scenarios; and explore the relationships between online content and other behaviours, actions and events. These steps form a necessary foundation to then identify and examine different opportunities for the responsible regulation of social media. The “Digital Wildfire” project undertaken by the authors of this paper advances this research agenda and in doing so seeks to make academic contributions and produce significant practical impacts.

## ACKNOWLEDGEMENTS

We would like to thank the journal editors and anonymous reviewers for their helpful comments on previous versions of this paper. We would also like to thank the Economic and Social Research Council (ESRC), the funders of the ‘Digital Wildfire’ study under Grant Number ES/L013398/1.

## REFERENCES

ADAM, A. Computer ethics in a different voice, Information and Organization. 11, 4 (2001), 235–261.

ADLER, M. and ZIGLIO, E. (Eds.) Gazing into the oracle: the Delphi method and its application to social policy and public health. Jessica Kingsley Publishers. (1996).

ALBERT, R. and BARABASI, A.L. Statistical Mechanics of Complex Networks. Reviews of Modern Physics, 74,1 (2001), 47-97.

ALLPORT, F.H. and LEPKIN, M. Wartime rumors of waste and special privilege: Why some people believe them, Journal of Abnormal and Social Psychology. 40 (1945), 3-36.

ARAL, S., MUCHNIK, L. and SUNDARARAJAN, A. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks, Proc. Nat, Acad. Sci. (PNAS). 106, 51 (2009), 21544-21549.

AWAN, I. Islamophobia and Twitter: A Typology of Online Hate Against Muslims on Social Media. Policy & Internet 6, 2 (2014) 133-150.

BACKSTROM, L., KLEINBERG, J., LEE, L. and DANESCU-NICULESCU-MIZIL, C. Characterizing and curating conversation threads: expansion, focus, volume, re-entry, Paper presented at the Proceedings of the sixth ACM international conference on Web search and data mining, Rome, Italy. (2013).

BAKER, S.A. From the criminal crowd to the “mediated crowd”: the impact of social media on the 2011 English riots, Safer communities. 11, 1 (2012), 40-49.

BANDARI, R., ASUR, S. and HUBERMAN, B.A. The Pulse of News in Social Media: Forecasting Popularity, CoRR. abs/1202.0332, (2012).

BAO, Y., Yi, C., Xue, Y. and Dong, Y. A new rumor propagation model and control strategy on social networks. In Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM '13). ACM, New York, NY, USA, (2013), 1472-1473. DOI: http://doi.acm.org/10.1145/2492517.2492599.

BARRAT, B., BARTHELEMY, M. and VESPIGNANI, A. Dynamical Processes on Complex Networks. Cambridge University Press, Cambridge, (2008).

BASSELL, L. Media and the Riots - A Call For Action, Citizen Journalism Educational Trust and The Latest.com. (2012). Available at http://www.the-latest.com/riots-and-media-report.

BBC NEWS. China punishes 197 over stock market and Tianjin 'rumours', bbc.co.uk/news 30 Aug 2015 (2015). Retrieved 30 August 2015 from http://www.bbc.co.uk/news/world-asia-china-34104114

BLANCHARD, B., LI, H. and CARSTEN, P. China threatens tough punishment for online rumour spreading, reuters.com 9 Sep 2013 (2013). Retrieved 17 August 2015 from http://www.reuters.com/article/2013/09/09/us-china-internet-idUSBRE9880CQ20130909.

BRUNS, A. Blogs, Wikipedia, Second Life, and beyond: From production to produsage (Digital Formations Vol. 45). Peter Lang, (2008).

BURNAP, P. and WILLIAMS, M. Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making, Policy & Internet 7,2 (2015).

BURNAP, P., RANA, O., AVIS, N., WILLIAMS, M.L., HOUSLEY, W., EDWARDS, A., MORGAN, J., and SLOAN, L. Detecting Tension in Online Communities with Computational Twitter Analysis, Technological Forecasting and Social Change. (2013). Retrieved 17 August 2015 from http://www.sciencedirect.com/science/article/pii/S0040162513000899.

BURNAP, P., RANA, O., WILLIAMS, M., HOUSLEY, W., EDWARDS, A., MORGAN, J, SLOAN, L. and CONEJERO, J. COSMOS: Towards an Integrated and Scalable Service for Analyzing Social Media on Demand, International Journal of Parallel, Emergent and Distributed Systems (IJPEDS). 30,2 (2014a).

BURNAP, P., WILLIAMS, M.L, SLOAN, L., RANA, O., HOUSLEY, W., EDWARDS, A., KNIGHT, V., PROCTER, R. and Voss, A. Tweeting the Terror: Modelling the Social Media Reaction to the Woolwich Terrorist Attack, Social Network Analysis and Mining. 4,1 (2014b).

BYNUM, T.W. and ROGERSON, S. Computer Ethics and Professional Responsibility: Introductory Text and Readings. Wiley Blackwell, New York (2003).

CASILLI, A.A. and TUBARO, P. Why net censorship in times of political unrest results in more violent uprisings: A social simulation experiment on the UK riots, SSRN eLibrary. 14, (2011).

CHEI, S.L. and LONG, M. News sharing in social media: The effect of gratifications and prior experience, Computers in Human Behavior. 28, 2 (2011), 331-339. DOI:http://dx.doi.org/10.1016/j.chb.2011.10.002.

CHIERICHETTI, F., LATTANZI, S. and PANCONESI, A. Rumour spreading in social networks. Automata, Languages and Programming. Springer, Berlin Heidelberg. (2009), 375-386.

COHEN, S. Folk devils and moral panics the creation of the Mods and Rockers. Paladin, London (1973).

COULTHARD M. An introduction to discourse analysis. London, Longman, (1977).F

CROWN PROSECUTION SERVICE. Guidelines on prosecuting cases involving communications sent via social media, CPS. (2013) Retrieved 24 March 2015 from http://www.cps.gov.uk/legal/a_to_c/communications_sent_via_social_media/.

DAHLGREN, P. Political participation via the web: structural and subjective contingencies, Interactions: Studies in Communication and Culture. 5,3 (2014), 255-269. DOI: http://dx.doi.org/10.1386/iscc.5.3.255_1.

DAVISON, J. Amateur online sleuthing: does it do more harm than good? CBC News 19 April 2013 (2013). Retrieved 12 August 2015 from http://www.cbc.ca/news/technology/amateur-online-sleuthing-does-it-do-more-harm-than-good-1.1412039 .

DAWKINS, R. Memes: the new replicators, In The Selfish Gene (2nd ed.). Oxford University Press, Oxford. (1989).

DECHUN L. and CHEN, X. Rumor Propagation in Online Social Networks Like Twitter -- A Simulation Study. Third International Conference on Multimedia Information Networking and Security (MINES) 4-6 Nov. 2011. (2011), 278,282.

DERCZYNSKI, L., BONTCHEVA, K., LUKASIK, M., DECLERCK, T., SCHARL, A., GEORGIEV, G., PROCTER, R., TOLMIE, P., ZUBIAGA, A., & LIAKATA, M. PHEME: Computing Veracity—the Fourth Challenge of Big Social Data (2015). Retrieved 16 August 2015 from http://derczynski.com/sheffield/papers/pheme-eswc-pn.pdf .

DIJK, T. van (Ed.) Handbook of Discourse Analysis, Vol 3: Discourse and Dialogue. London, Academic. (1985).

DINGWALL, R. Contemporary legends, rumours and collective behaviour: Some neglected resources for medical sociology? Sociology of Health & Illness. 23,2 (2001), 180-202.

DOERR, B., FOUZ, M and FRIEDRICH, T. Why rumors spread so quickly in social networks, Communications of the ACM 55.6 (2012), 70-75.

DOSHI, S. Building a safer twitter, twitter.com 2 Dec 2014 (2014). Retrieved 15 Jan 2015 from https://blog.twitter.com/2014/building-a-safer-twitter.

EDWARDS, A., HOUSLEY, W., WILLIAMS, Matthew, SLOAN, L. and WILLIAMS, Malcolm. Digital social research, social media and the sociological imagination: surrogacy, augmentation and re-orientation, International Journal of Social Research Methodology 16,3 (2013), 245-260. DOI:10.1080/13645579.2013.774185.

FLORIDI, L. (2010). Information ethics. In L. FLORIDI (Ed.) The Cambridge Handbook of Information and Computer Ethics. Cambridge University Press, Cambridge, (2010), 77–97.

GAGLIARDONE, I., GAL, D., ALVES, T. and MARTINEZ, G. Countering online hate speech. UNESCO series on internet freedom. United Nations Educational, Scientific and Cultural Organisation, Paris, (2015).

GOFFMAN, E. Forms of Talk. University of Pennsylvania Press, Pennsylvania (1981).

GIL DE ZÚÑIGA, H., JUNG, N. and VALENZUELA, S. Social Media Use for News and Individuals' Social Capital, Civic Engagement and Political Participation, Journal of Computer-Mediated Communication. 17 (2012), 319–336. DOI:10.1111/j.1083-6101.2012.01574.

GOLDENBERG J. LIBAI B., and MULLER E. Talk of the network: A complex systems look at the underlying process of word-of-mouth, Marketing Letters. 3, 12, (2001), 211–223.

GUILLE A. and HACID, H. A predictive model for the temporal dynamics of information diffusion in online social networks, Paper presented at the 21st International conference companion on World Wide Web, Lyon, France. (2012).

GULF CENTRE FOR HUMAN RIGHTS. Qatar: new cyber crime law poses real threat to freedom of expression, gc4hr.org 17 Sep 2014 (2014). Retrieved 17 August 2015 from http://www.gc4hr.org/news/view/747.

GUPTA, A., LAMBA, H., KUMARAGURU, P. and JOSHI, A. Faking Sandy: Characterising and Identifying fake images on Twitter during Hurricane Sandy, Second International Workshop on Privacy and Security on Social Media (PSOSM) May 2013, (2013) Retrieved 17 August 2015 from http://ebiquity.umbc.edu/paper/html/id/623/Faking-Sandy-Characterizing-and-Identifying-Fake-Images-on-Twitter-during-Hurricane-Sandy .

HARDAKER, C. Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions, Journal of Politeness Research 6,2 (2010), 215-242.

HER MAJESTY'S INSPECTORATE OF CONSTABULARY. The Rules of Engagement: A Review of the August 2011 Disorders. HMIC, London, (2011a).

HER MAJESTY'S INSPECTORATE OF CONSTABULARY. Policing Public Order: An overview and review of progress against the recommendations of Adapting to Protest and Nurturing the British Model of Policing. HMIC, London, (2011b).

HOUSLEY, W., PROCTER, R., EDWARDS, A., BURNAP, P., WILLIAMS, M., SLOAN, L., RANA, O., MORGAN, J., VOSS, A. and GREENHILL, G. Big and broad social data and the sociological imagination: A collaborative response, Big Data & Society. 1, 2 (2014).

HOUSE OF LORDS. Social Media and Criminal Offences: 1st report of Session 2014-2015. The Stationery Office Limited, London, (2014). Retrieved 20 December 2015: http://www.publications.parliament.uk/pa/ld201415/ldselect/ldcomuni/37/3702.htm .

INNES, M. Signal crimes and signal disorders: notes on deviance as communicative action, British Journal of Sociology. 55, 3 (2004), 335–355.

JAVA, A., SONG X. FININ, T. and TSENG, B. Why we twitter: Understanding microblogging usage and communities, In Proc. ACM 9th WebKDD and 1st SNA-KDD Workshop on Web Mining and Social Network Analysis. (2007), 56-65.

JOHNSON, D.G. Computer Ethics (1st edition). Prentice Hall, Upper Saddle River, New Jersey, (1985).

KJØLBERG, K.A.L. The notion of 'responsible development' in new approaches to governance of nanosciences and nanotechnologies Doctoral dissertation, The University of Bergen. (2010).

KOSTKA, J., OSWALD, Y.A. and WATTENHOFER, R. Word of mouth: Rumor dissemination in social networks. Structural Information and Communication Complexity. Springer, Berlin Heidelberg. (2008), 185-196.

KWON, S. CHA, M., JUNG, K., CHEN, W. and WANG, Y. Prominent Features of Rumor Propagation in Online Social Media, 13th International Conference on Data Mining (ICDM), 2013, IEEE. (2013), 1103-1108.

LEWIS, P., NEWBURN, T., TAYLOR, M., MCGILLIVRAY, C., GREENHILL, A., FRAYMAN, H. and PROCTER, R. Reading the Riots: Investigating England's summer of disorder. (2011). Retrieved 24 March 2015 from http://www.guardian.co.uk/uk/series/reading-the-riots .

LOADER, B.D. and MERCEA, D. (Eds.) Social media and democracy. Routledge, London and New York, (2012).

LOTAN, G., GRAEFF, E., ANANNY, M., GAFFNEY, D., PEARCE, I. and Boyd, D. The Revolutions Were Tweeted: Information Flows During the 2011 Tunisian and Egyptian Revolutions, International Journal of Communication 5 (Special Issue) (2011), 1375-1405.

LUCKERSON, V. Fear, misinformation and social media complicate ebola fight, Time 8 Oct 2014. (2014), Retrieved 24 March 2015 from http://time.com/3479254/ebola-social-media/ .

LUPTON, D. Digital Sociology. Routledge, London (2015).

MACSKASSY, S. and MICHELSON, M. Why do people retweet? antihomophily wins the day, In: International Conference on Weblogs and Social Media (ICWSM). (2011).

McCOSKER, A. Trolling as provocation YouTube's agonistic publics. Convergence: The International Journal of Research into New Media Technologies. 20,2 (2014), 201-217.

McENERY, T., McGLASHAN, M and LOVE, R. Press and media reaction to ideologically inspired murder: the case of Lee Rigby, Discourse and Communication. 9,2 (2015), 237-259.

MENDOZA, M., POBLETE, B. and CASTILLO, C. Twitter under Crisis: Can We Trust What We RT? In 1st Workshop on Social Media Analytics (SOMA '10). ACM Press, Washington, D.C. (2010).

MILLER, B. UK petition calls on Twitter to tackle abuse after Caroline Criado-Perez subjected to violent tweets, abc.net 27 July 2013 (2013). Retrieved 12 Jan 2015 from http://www.abc.net.au/news/2013-07-29/thousands-sign-petition-to-stop-abusive-tweets/4849780.

MORRIS, S. Contagion, Review of Economic Studies. 67,1 (2000), 57–78.

MOSSBERGER, K., TOLBERT, C.J., and MCNEAL, R.S. Digital Citizenship: The Internet, Society and Participation. MIT Press, Massachusetts. (2008).

MUNSON, L. India strikes down controversial "Section 66A" social media policing law, nakedsecurity.com 25 March 2015 (2015). Retrieved 17 August from https://nakedsecurity.sophos.com/2015/03/25/india-strikes-down-controversial-section-66a-social-media-policing-law/.

MURTHY, D. Towards a sociological understanding of social media: theorizing Twitter, Sociology. 46, 6 (2012a), 1059-1073. DOI: 10.1177/0038038511422553.

MURTHY, D. Twitter: Social Communication in the Twitter age. Polity Press, Cambridge, UK. (2012b).

NEKOVEE M., MORENO, Y., BIANCONI, G. and MARSILI, M. Theory of rumour spreading in complex social networks, Physica A: Statistical Mechanics and its Applications. 374, 1 (2007), 457-470. DOI:http://dx.doi.org/10.1016/j.physa.2006.07.017.

NEWMAN, M.E.J. and PARK, J. Why social networks are different from other types of networks, Physical Review E. 68, 3 (2003),036122.

OWEN, R., MACNAGHTEN, P. and STILGOE, J. Responsible research and innovation: From science in society to science for society, with society, Science and Public Policy. 39 (2012), 751–760.

PICKLES, N. Safer Internet Day: protecting the global town square of Twitter, The Guardian 9 Feb 2016, (2016). Retrieved 9 Feb from http://www.theguardian.com/technology/2016/feb/08/twitter-safer-internet-day-nick-pickles-online-diversity?CMP=share_btn_tw .

PROCTER, R., VIS, F. and VOSS, A. Reading the riots on Twitter: methodological innovation for the analysis of big data, International Journal of Social Research Methodology. 16, 3 (2013a), 197-214. DOI:10.1080/13645579.2013.774172.

PROCTER, R., CRUMP, J., KARSTEDT, S., VOSS, A. and CANTIJOCH, M. Reading the riots: what were the police doing on Twitter? Policing and Society. 23, 4 (2013b), 1-24. DOI:10.1080/10439463.2013.780223.

RATKIEWICZ, J., CONOVER, M., MEISS, M., GONÇALVES, B., PATIL, S., FLAMMINI, A. and MENCZER, F. Truthy: mapping the spread of astroturf in microblog streams, in Proceedings of the 20th international conference companion on World wide web, ACM. (2011), 249-252.

ROCO, M.C., HARTHORN, B., GUSTON, D. and SHAPIRA, P. Innovative and responsible governance of nanotechnology for societal development, J. of Nanoparticle Research. 13 (2011), 3557–3590.

ROMERO D.M., MEEDER, B. and KLEINBERG, J. Differences in the mechanics of information diffusion across topics: Idioms, political hashtags and complex contagion on Twitter, Proc.Intl. Conf. on World Wide Web (WWW). (2011).

RUPPERT, E., LAW, J. and SAVAGE, M. Reassembling social science methods: the challenge of digital devices, Theory, Culture & Society. 30,4 (2013), 22-46.

SACKS, H., SCHEGLOFF, E. A., and JEFFERSON, G. A simplest systematics for the organization of turn-taking for conversation, Language (1974), 696-735.

SHIBUTANI, T. Improvised news: A sociological study of rumor. Bobb-Merrill, Indianapolis. (1966).

STAHL B.C. Morality, Ethics and Reflection: A categorisation of normative research in IS research, Journal of the Association for Information Systems. 13, 8 (2012), 636–656. Retrieved 11 Jan 2015 from http://aisel.aisnet.org/jais/vol13/iss8/1/.

STAHL B.C., EDEN, G., JIROTKA, M. and COECKELBERGH, M. From Computer Ethics to Responsible Research and Innovation in ICT: The transition of reference discourses informing ethics-related research in information systems, Information & Management. 51, 6 (2014), 810-818. DOI:10.1016/j.im.2014.01.001 .

STILGOE, J. OWEN, R. and MACNAGHTEN, P. Developing a framework for responsible innovation, Research Policy. 42, 9 (2013), 1568-1680. DOI:http://dx.doi.org/10.1016/j.respol.2013.05.008 .

SUH, B., HONG, L., PIROLLI, P., and CHI, E. Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network, IEEE Second Conference on SocialCom. (2011).

TOLMIE, P., PROCTER, R., ROUNCEFIELD, M., LIAKATA, M., and ZUBIAGA, A. (2015). Microblog Analysis as a Programme of Work. Submitted to ACM Transactions on Computer-Human Interaction. Available from *arXiv preprint arXiv:1511.03193*.

TONKIN, E., PFEIFFER, H.D. and TOURTE, G. Twitter, information sharing and the London riots? Bulletin of the American Society for Information Science and Technology. 38, 2 (2012), 49-57.

TROTTIER, D. Social Media as Surveillance. Ashgate, Surrey, England. (2012).

TRENHOLM, R. Cameron considers blocking Twitter, Facebook, BBM after riots, CNET 11 August 2011. (2011) Retrieved March 26, 2015 from http://www.cnet.com/uk/news/cameron-considers-blocking-twitter-facebook-bbm-after-riots/ .

TSUR, O. and RAPPOPORT, A. What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities, Paper presented at the Proceedings of the fifth ACM international conference on Web search and data mining, Seattle, Washington, USA. (2012).

TWEED, P. Lord McAlpine and the high cost of tweeting gossip. The Guardian 27 Nov 2012 (2012). Retrieved 21 March 2015 from http://www.theguardian.com/law/2012/nov/27/lord-mcalpine-twitter-libel .

TUYSUZ, G. Turkey blocks social media websites, CNN.com 6 April 2015 (2015). Retrieved 17 August 2015 from http://edition.cnn.com/2015/04/06/world/turkey-social-media-blocked/.

UK SAFER INTERNET CENTRE. Safer Internet day 2015, saferinternet.org 10 Feb 2015 (2015). Retrieved 24 March 2015 from http://www.saferinternet.org.uk/safer-internet-day/2015 .

WAM. Harrassement of women on Twitter We're on it!, Women Action and the Media 6 Nov 2014. (2014) Retrieved 20 Jan 2015 from http://www.womenactionmedia.org/2014/11/06/harassment-of-women-on-twitter-were-on-it/.

WATTS, D. and DODDS, P. Threshold models of social influence, in HEDSTROM, P. and BEARMAN, P.S (Eds), Oxford Handbook of Analytical Sociology, Oxford University Press, Oxford. (2009), 475-497.

WEBB, H., JIROTKA, M., CARSTEN STAHL, B., HOUSLEY, W., EDWARDS, A., WILLIAMS, M., PROCTER, R., RANA, O. and BURNAP, P. Digital wildfires: hyper-connectivity, havoc and a global ethos to govern social media, Computers and Society 45,3 (2015), 193-201. Retrieved 20 Oct 2015 from http://www.dmu.ac.uk/documents/research-documents/technology/ccsr/20-years-of-ethicomp-si.pdf.

WENDLING, C., RADISCH, J and JACOBZONE, S. The Use of Social Media in Risk and Crisis Communication, OECD Working Papers on Public Governance, No. 24, OECD Publishing (2013). DOI: http://dx.doi.org/10.1787/5k3v01fskp9s-en .

WILKINS L. Social Deviance. Tavistock Publications, London. (1967).

WILLIAMS, M. L. and BURNAP, P. Cyberhate on social media in the aftermath of Woolwich: A case study in computational criminology and big data. British Journal of Criminology 56, 2 (2015), 211-238.

WILLIAMS, M.L., EDWARDS, A., HOUSLEY, W., BURNAP, P., RANA, O., AVIS, N., MORGAN, J. and SLOAN, L. Policing Cyber-Neighbourhoods: Tension Monitoring and Social Media Networks, Policing & Society. 24, 4 (2013), 461-481.

WORLD ECONOMIC FORUM. Digital Wildfires in a hyperconnected world. Global Risks Report, World Economic Forum. (2013). Retrieved 20 Nov 2014 http://reports.weforum.org/global-risks-2013/risk-case-1/digital-wildfires-in-a-hyperconnected-world/ .

YANG, J. and COUNTS, S. Predicting the Speed, Scale, and Range of Information Diffusion in Twitter, International Conference on Weblogs and Social Media (ICWSM). (2010).

ZAMAN, T, HERBRICH, R, VAN GAEL, J. and STERN, D. Predicting information spreading in Twitter, Workshop on Computational Social Science and the Wisdom of Crowds (NIPS). (2010).

ZAMAN, T., FOX, E. and BRADLOW, E. A Bayesian Approach for Predicting the Popularity of Tweets, CoRR. (2013).

ZUBIAGA, A., LIAKATA, M., PROCTER, R., BONTCHEVA, K. and TOLMIE, P. Towards Detecting Rumours in Social Media. Proceedings of the AAAI Workshop on AI for Cities (2015a)..

ZUBIAGA, A., SPINA, D., MARTÍNEZ, R. and FRESNO, V. Real-time classification of Twitter trends. Journal of the Association for Information Science and Technology, 66 (2015b) 462–473.

ZUBIAGA, A., LIAKATA, M., PROCTER, R., TOLMIE, P. and WONG SAK HOI, G. Analysing How People Orient to and Spread Rumours in Social Media by Looking at Conversational Threads. PLOS One (2016). Available from *arXiv preprint arXiv:1511.07487*.