Proceedings of the Tenth International AAAI Conference on
Web and Social Media (ICWSM 2016)

Mining Pro-ISIS Radicalisation Signals from Social Media Users

Matthew Rowe
School of Computing and Communications
Lancaster University
Lancaster, United Kingdeom
m.rowe@lancaster.ac.uk

Hassan Saif
Knowledge Media Institute
The Open University
Milton Keynes, United Kingdeom
h.saif@open.ac.uk

Abstract

The emergence and actions of the so-called Islamic State of Iraq and the Levant (ISIL/ISIS) has received widespread news coverage across the World, largely due to their capture of large swathes of land across Syria and Iraq, and the publishing of execution and propaganda videos. Enticed by such material published on social media and attracted to the cause of ISIS, there have been numerous reports of individuals from European countries (the United Kingdom and France in particular) moving to Syria and joining ISIS. In this paper our aim to understand what happens to Europe-based Twitter users before, during, and after they exhibit pro-ISIS behaviour (i.e. using pro-ISIS terms, sharing content from pro-ISIS accounts), characterising such behaviour as radicalisation signals. We adopt a data-mining oriented approach to computationally determine time points of activation (i.e. when users begin to adopt pro-ISIS behaviour), characterise divergent behaviour (both lexically and socially), and quantify influence dynamics as pro-ISIS terms are adopted. Our findings show that: (i) of 154K users examined only 727 exhibited signs of pro-ISIS behaviour and the vast majority of those 727 users became *activated* with such behaviour during the summer of 2014 when ISIS shared many beheading videos online; (ii) users exhibit significant behaviour divergence around the time of their activation, and; (iii) social homophily has a strong bearing on the diffusion process of pro-ISIS terms through Twitter.

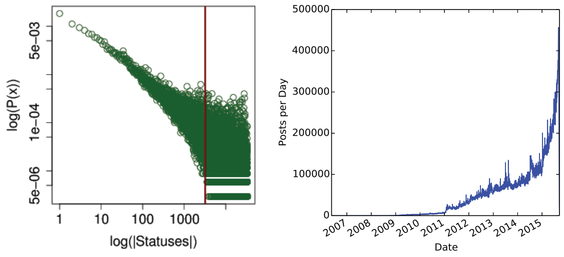
Introduction

The Arab Spring of 2011 brought about widespread protests in the Middle-East and led to democratic elections taking place in several countries (e.g. Egypt), however protests in many countries were not as successful at instigating democratic change. In particular, uprisings in Syria against the government of president Bashar al-Assad escalated into a civil war, originally fought between the Free Syrian Army and government forces. The instability that the conflict cultured in turn led to the so-called Islamic State of Iraq and the Levant (ISIL/ISIS) seizing control of vast swathes of land in Syria and northern Iraq throughout 2013, and the instigation of Sharia law throughout those areas. Since 2013, ISIS have been proactive in using online propaganda to highlight their work, to recruit Westerners - in particular Muslims from European countries to join them in Syria - and to carry out terrorist activities in western countries. There have been numerous reports of people from European countries, in particular the United Kingdom, France and Belgium, moving to Syria to join ISIS: in essence going through a process of *radicalisation* where their views become conflicted with the (Western) society in which they are residing.

Recent research has sought to understand ISIS's social media presence (Bazan, Saad, and Chamoun 2015; Winter 2015; Berger and Morgan 2015; Klausen 2015), the process of online radicalisation (Bermingham et al. 2009; Edwards and Gribbon 2013; Torok 2013), and the various stages that online radicalisation is comprised of (Bartlett and Miller 2012; King and Taylor 2011; Berger 2015; Hall 2015). However, what is not currently understood is what happens to social media users *before* they adopt pro-ISIS behaviour (i.e. using pro-ISIS language, sharing content from pro-ISIS accounts) and how they *develop* into this state. Understanding this could not only pave the way to *detecting* if a user is *likely* to adopt a pro-ISIS stance, but also understanding the context under which this occurs so that counter-narratives to radicalisation can be devised - something that researchers from studying radicalisation have noted governments' omission of.

Motivated by this dearth in understanding, we sought to investigate the following research questions: **RQ1**: *How can we detect when a user has adopted a pro-ISIS stance (i.e. is exhibiting radicalised behaviour on social media)?* **RQ2**: *What happens to Twitter users before they exhibit radicalised behaviour, and also after such exhibition?* And **RQ3**: *What influences users to adopt pro-ISIS language?* In this paper we describe our methodology and findings in the pursuit of the above questions. Using a data-mining oriented approach, we were able to quantify signals of 'radicalisation' based on users adopting known pro-ISIS terms and sharing (i.e. *retweeting*) content from suspended and known pro-ISIS Twitter accounts. By identifying when such users became *activated* with such radicalised behaviour, we were able to examine how users' behaviour (in terms of language used, and social interactions) before, during, and after their activation changed. Furthermore, by treating the activation of users as a diffusion process, we found that users adopted pro-ISIS terms from users with whom they had high levels of *social homophily*, indicating the presence of common sub-communities of users from whom radicalised content is shared.

We have structured this paper as follows: section 2 describes the data collected from Twitter for our work and its characteristics. The following section assesses the related work in the areas of radicalisation studies and examining pro-ISIS users online, after which we define our '*radicalisation hypotheses*' that we use to identify a user as exhibiting pro-ISIS behaviour. We follow this by analysing the activation points of users and what happens to them before, during and after activation. We then investigate the pathways that users go through before they become activated and what factors *influence* them to adopt pro-ISIS terms from other users (operationalised using a general threshold diffusion model), and then finally conclude the paper with a discussion of our findings and their implications on studies of online radicalisation.



(a) Status count per user (b) Tweets per day

Figure 1: The distribution of per-user status counts is heavy tailed (Figure 1(a) - the red line indicates the 3,200 threshold for which we can collect Tweets for from the twitter API, while the collected dataset of Tweets from 154K users within Europe exhibits a large surge in activity towards the end (Figure 1(b)).

Data Collection and Initial Analysis

Our first task was to collect a dataset of Twitter users, together with their posts, who resided in Europe and that contained pro-ISIS, anti-ISIS, and neutral users. As a starting point, we were provided with the Twitter user ids of 652 users that featured in prior work by O'Callaghan et al (O'Callaghan et al. 2014) - these were used as *seed* accounts collected from Twitter lists pertaining to the Syria conflict, so would contain a mix of pro-ISIS, anti-ISIS, and neutral users. We began by checking to see which of these users were still active, not deleted, and had their timeline visible: we found that 512 were still available for use.[1] From those 512 users, we then collected the followers of those users, resulting in a collection of 2.4M users. We pruned this set of users down to only those users who described themselves as being based within Europe: for this we used a gazetteer of European location names and countries, and a basic string matching comparison between each user's biography location and the gazetteer to ensure strict matching, only using location names that are unique to European countries or where the user had defined themselves explicitly within a country. After performing this *filtering* step we were left with a pruned set of 153,947 ($\sim$ 154K) users who resided in Europe. These users form the basis for our analysis.

Given that we were interested in studying the behaviour of users *prior* to them exhibiting radicalised behaviour, we needed to gauge the degree to which we could gather users' timelines - given that the Twitter REST API imposes a limit of what can be retrieved from a users' timeline to the most recent 3,200 Tweets. Therefore, we derived the *status count* distribution across the 154K users and plotted this - as shown in Figure 1(a), with the 3,200 status threshold indicated by the solid red line as the maximum number of Tweets that we can collect from a present point in time going back for a given user. Based on this threshold, for 97% of users we can collect the *full timeline*, and thus all of their Tweets, while for all users in our sample we can collect over half of their timeline. We proceeded with collecting the maximum 3,200 tweets per user which resulted in a dataset containing 104,347,769 ($\sim$ 104M) tweets.[2] Further inspection of the 104M Tweets collected from 154K users reveals an increased activity trend as time increases (Figure 1(b)). The language distribution of Tweets in the dataset is found to be mainly English (43%) and Arabic (41%), with Dutch and then Spanish: in this paper we only process Tweets that have been identified as English and Arabic, as the authors are fluent in these languages.

Related Work

ISIS on Social Media

The rise of ISIS and its use of social media, arguably in an effective way, has caught several Western nations' governments off guard - especially when one considers the counter narrative (or lack of) to ISIS's propaganda online. That said, researchers from the areas of counter-terrorism and cyber-security have begun to examine this space and to understand ISIS's social media presence and actions. For instance, Bazan et al. (Bazan, Saad, and Chamoun 2015) examined the '*information warfare*' performed by ISIS. The authors defined a typology of warfare actions using content platforms and apps, performing hacking, and propaganda techniques (e.g. mujatweets); together with ISIS's three-fold strategic objectives: (i) to build and support an audience online, (ii) frame politico-military objectives and explain them to the public, and (iii) market the caliphate as a strategic priority and attractive proposition. In a similar vein, Winter (Winter 2015) performed an in-depth systematic analysis of the media output from Islamic State's Central Media Command, finding that the majority of their content seeks to highlight their aim to establish a utopia in their intended caliphate. Winter argued that conveying the appearance of this utopia via social media is essential to ISIS, and allows

[1] Note that we do not know what happened to the remaining 140 users as we can no longer access their data.

[2] N.b. Should the paper be accepted we will be sharing the Tweet IDs with researchers upon request - as per Twitter's terms and conditions.

the audience to see how it would be to live there (and thus potentially entice them to the cause) by covering such topics as justice, governance, and economic activity.

Berger and Morgan (Berger and Morgan 2015) carried out social network analysis of manually collected pro-ISIS Twitter accounts, and estimated that there are as many as 90K ISIS supporters on Twitter (as of summer 2015). That said, the authors found limited numbers of users residing in Europe (i.e. <150 in each of UK, France and Belgium). The authors found that pro-ISIS supporters could be predicted from the terms used in their profile descriptions: with terms such as succession, linger, Islamic State, Caliphate State or In Iraq all being prominent - putting this information to use Berger and Morgan achieved 94% accuracy in differentiating pro and anti-ISIS supporters. Similar, Klausen (Klausen 2015) examined the Twitter accounts of known 59 Western-origin fighters in Syria. The author found that these accounts followed specific common operative accounts in Syria (that release ISIS propaganda) - from Islamic State's Central Media Command (Winter 2015) - and also accounts from UK-based banned organisation Al Muhajiroun.

Online Radicalisation

Understanding the process of *radicalisation* through the Internet and what motivates individuals to adopt such behaviour has been the focus of several studies. One of the first pieces of work in this space was that of Bermingham et al. (Bermingham et al. 2009) who looked at the language used by users who are subscribed to Jihadi YouTube video groups, finding common patterns of language in terms of the top-terms used within such groups - often focussing on discussions of religion and not necessarily about influence or recruitment to a given ideology. Edwards and Gribbon (Edwards and Gribbon 2013) investigated Internet Radicalisation in the UK by speaking with convicted terrorists and those known to have been radicalised online. One of the salient findings of their work that the the process of radicalisation is being '*increasingly covert*' where individuals are not attending mosques to discuss radical views, but are instead turning to the Internet to find information inline with extreme beliefs (i.e. several radicalised people had viewed or shared beheading videos). The authors point out that 2 of study's subjects actively sought to *radicalise* other people online, where the subjects explained that *journeys* of radicalisation differ between people and thus occur at varying rates, however a common *snowball* effect in consumption and sharing of extremist material was evident for several subjects. In order to understand the power structures in online groups and communities that aim to radicalise members, Torok (Torok 2013) took a grounded theory approach. This involved examining 10 social media groups on Facebook to perform a qualitative analysis of what was being discussed, by whom, and the observed power structure that existed in such communities. Torok found that key *influential* members' (i.e. *elders*) discourse in this community carried more weight than others, and had a greater potential for conversion to a given viewpoint.

Examination of the *process* of radicalisation and the various stages and factors that it contains was undertaken by Bartlett and Miller (Bartlett and Miller 2012) using primary data obtained from court reports and focus groups and interviews. Bartlett and Miller found common *signifiers* of movement towards radicalisation such as the distribution of jihad videos, clashing with existing mosque authorities, and engagement in literature defining what a '*kafir*' is. Similarly, King and Taylor (King and Taylor 2011) presented an an overview of five radicalisation '*pathway*' models that document an individual's journey, coding an individual as having become radicalised if he/she advocates or looks to partake in violence and/or terrorism: common radicalisation signifiers were the state of blaming the West (e.g. the US) for the ills of a given group (often brought together under a common focus - e.g. religion, political niche).

Berger (Berger 2015) defined the online radicalisation/recruitment to ISIS as being a five-part process: discovery of a potential recruit, creation of a micro-community (where ISIS supporters ingratiate themselves to the candidate), isolation of the candidate (via severing of ties with family and friends), use of private communities (to discuss travel/logistics), and then the encouragement of action (either travelling to Syria/Iraq or performing a resident country act). Berger's framework is more prescribed than the process elicited by Hall (Hall 2015) when examining the recruitment of Canadian supporters to the ISIS cause, instead focussing more on the initial *tempting* of sympathisers via propaganda using existing recruited fighters. In particular, Hall found that Islamic State's digital magazine *Dabiq* focussed on legitimising the caliphate and its normalisation - much in line with the findings of Winter (Winter 2015).

Our extensive examination of related works in both the areas of ISIS on social media and studies of online radicalisation demonstrates how advanced ISIS are in using social media to spread their message and catch the attention of potential recruits. In this paper we focus on mining *radicalisation signals* from users (i.e. we do not explicitly say that an individual has become *radicalised*) based on changes in their behaviour. This novel approach means that we fill two clear gaps in the related work: firstly, we provide an inspection of users' development over time before, during and after their *activation* with pro-ISIS behaviour, and; secondly, we determine under what conditions users are *influenced* to adopt pro-ISIS terms in their language.

Identifying Signals of Radicalisation

In order to understand when a user has shifted to a pro-ISIS position, and to investigate **RQ1**: *How can we detect when a user has adopted a pro-ISIS stance (i.e. is exhibiting radicalised behaviour on social media)?*, requires inspection of their behaviour over time for critical points of *activation*. Here, we treat such *activation* a binary switch such that a user either *exhibits* pro-ISIS behaviour or does not - in essence emitting a signal of radicalisation. Based on our study of the literature above, salient properties of radicalisation signals are often the sharing of pro-ISIS content (Bartlett and Miller 2012) and using pro-ISIS language/rhetoric (King and Taylor 2011) . Using such assumptions we posit the following two hypotheses that we use to identify users in our dataset as being pro-ISIS, or not:

- **H1 - Sharing Incitement Material:** The user shares tweets from either known pro-ISIS accounts or accounts that have been suspended for supporting ISIS.[3] Here we are focusing on the user's action of *passing on* extremist material and hence the role of diffusion.
- **H2 - Using Extremist Language:** The user adds certain *keywords* to their tweets that are synonymous with anti-Western and pro-ISIS rhetoric. Here we used a *lexicon* of terms identified from our own review of the collected users' tweets, and also from prior work (Bermingham et al. 2009; Berger and Morgan 2015) and suggested anti-Western rhetoric from (King and Taylor 2011). We define a user as using *pro-ISIS* language if they use pro-ISIS terms (from our lexicon) more than anti-ISIS terms, and use a given pro-ISIS term from the lexicon more than 5 times.

We applied the above hypotheses over our sample of 154K users and derived the following: 508 users in the set of H1 users (who had shared content from known pro-ISIS accounts or those suspended), 208 users in the set of H2 users, 727 users within the union of H1 and H2 users, and 64 users in the intersection of the sets of H1 and H2 users. In comparison with Berger and Morgan's work (Berger and Morgan 2015) these numbers are similar and reflect the *sparsity* in pro-ISIS users in the sample - Berger and Morgan found <150 pro-ISIS accounts per European country.

Lexicon Validation

In order to apply H2 we manually constructed a lexicon consisting of both pro-ISIS and anti-ISIS terms in both English and Arabic - this was constructed following our review of the related work above and speaking with researchers from the domain of religious studies who have investigated online radicalisation. To validate the lexicon's terms and to ensure that they were placed within the correct group (pro/anti), we ran a small annotation exercise. Two raters who were fluent in both English and Arabic, and who originated from the Middle-East, manually labelled a sample of 2K Tweets: 1K contained pro-ISIS terms from our lexicon (500 Arabic, 500 English), and 1K contained anti-ISIS terms (500 Arabic, 500 English) - the two raters labelled each Tweet as either: pro-ISIS, anti-ISIS, or neutral. After labelling, we then calculated Feiss's κ (Fleiss, Levin, and Paik 2013) to gauge the *interrater agreement* between the raters - with a value of 0 indicating total disagreement and 1 indicating total agreement. Our κ results were as follows: 0.418 for English pro-ISIS term Tweets and 0.504 for Arabic pro-ISIS term Tweets, and 0.439 for English anti-ISIS term Tweets and 0.521 for Arabic anti-ISIS term Tweets. This resulted in overall an overall κ value of 0.509, and 0.415 for English language tweets and 0.593 for Arabic language tweets According to Fleiss's table (Fleiss, Levin, and Paik 2013) for interpreting the κ value, we have consistent agreement values in the interval $[0.4, 0.6]$ which is defined as '*fairly-good*'.

Table 1: Lexicon terms with their original allocation to either originally indicating pro or anti-ISIS with the proportion of tweets that they appear in (i.e. either pro or anti-ISIS) in the labelled 2K sample.

	Orig' Label	Pro Prop'	Anti Prop'
English Terms			
Apostate	Pro	**0.666**	0.333
Caliphate	Pro	**0.524**	0.476
Islamic State	Pro	0.221	**0.779**
Khilafah	Pro	**0.909**	0.091
Shirk	Pro	**1.000**	0.000
Ummah	Pro	**0.692**	0.308
Daesh	Anti	0.000	**1.000**
Isis	Anti	0.066	**0.934**
Isil	Anti	0.000	**1.000**
Arabic Terms			
دولة الخلَافة	Pro	0.200	**0.800**
الخليفة	Pro	0.500	0.500
الدولة الَاسلَامية	Pro	**0.524**	0.476
ارهَاب	Anti	0.083	**0.917**
ارهَابيّن	Anti	0.000	**1.000**
دَاعش	Anti	0.000	**1.000**

Having labelled tweets, we then examined the proportion of pro or anti-ISIS tweets that each term in the pro-ISIS and anti-ISIS lexicons appeared in (using the agreed upon label between the raters as the actual label of the Tweet - i.e. if both raters label a Tweet as pro-ISIS then we label it as such, otherwise we discard the Tweet from our analysis).[4]

Our results from this validation exercise are shown in Table 1, where we note the original labelling of the term (either pro-ISIS or anti-ISIS) and the resultant proportions of pro or anti-ISIS Tweets that the term appears in. For two terms, Islamic State and دولة الخلَافة ('*The state of the Khilafat*') we originally labelled these terms as pro-ISIS, however they appear in more anti-ISIS tweets. Likewise, we labelled الخليفة (translates as '*Khilafah*') as pro-ISIS, however we found an equal proportion of pro and anti-ISIS tweets containing this term. Therefore we removed all these terms from our lexicon, and used the remaining pro and anti-ISIS terms as indicators of either pro or anti-ISIS sentiment - note that H2 looks at *all* terms used by a given user, and should she use more pro-ISIS than anti-ISIS terms then we assume that she is exhibiting a radicalisation signal.

Activation Points

We define users as becoming *activated* when they are classified as sharing radicalised content by either H1 or H2, from above. We begin by looking at *when* such activations occur and how such points may differ between the hypothe-

[3] We gathered these accounts from the online-hacking group anonymous and manually validated a sample of their retweeted content.

[4] We do not report the *neutral* class here, as we are concerned with the balance between pro and anti-ISIS appearance for our later diffusion experiments - given that neutral Tweets often act as *bridges* for information to spread following exposure.

Table 2: Significant events involving ISIS/ISIL and the West.

Date	Description
08-04-2013	ISIS expand into Syria
04-01-2014	Fallujah captured by ISIS
15-01-2014	ISIL retake Ar-Raqqah
01-05-2014	ISIS carry out public executions in Ar-Raqqah
09-06-2014	Mosul falls under ISIS control
02-09-2014	Hostage Steven Sotloff executed
13-09-2014	Hostage David Haines executed
22-09-2014	Hostage Samira Salih al-Nuaimi executed
03-10-2014	Hostage Alan Henning executed
07-10-2014	Abu Bakr al-Baghdadi injured in US air strike
16-10-2014	Hostage Peter Kassig executed
14-01-2015	Christopher Lee Cornell arrested for bomb plot
25-01-2015	Hostage Haruna Yukawa executed
31-01-2015	Hotage Kenji Goto executed
06-02-2015	Hostage Kayla Mueller killed in air strike
26-02-2015	Jihadi John is identified as Mohammed Emwazi
18-03-2015	ISIS responsible for Tunisia museum attack
15-05-2015	Abu Sayyaf killed by US special forces
30-06-2015	Alaa Saadeh arrested for attempts to aid ISIS
11-07-2015	Maher Meshaal killed in coalition air strike

ses. Figure 2(a) and figure 2(b) show the number of users who are activated on each day according to each hypothesis. We note that the span of activations of H1 users is shorter than H2 users - as the former requires sharing content from banned or pro-ISIS accounts, while the latter looks at the use of pro-ISIS terms. One thing that is immediately apparent from the plots is that there is a large surge in activity from May 2014 onwards - for both H1 and H2 activations. To investigate why this surge occurs, we identified a series of key events related to ISIS/ISIL from 2013 onwards - these are shown in Table 2. As noted, the increase in *activations* between May 2014 and November 2014 coincides with execution of 6 hostages by ISIS and the videos of these executions posted via social media. Although we cannot discern causation (of activation) from correlation here, there does appear to be an association between such information appearing in the public domain (of executions) and users either sharing pro-ISIS content (Figure 2(a)) or adopting pro-ISIS language (Figure 2(b)).

In order to examine whether there was a link between users sharing content from pro-ISIS accounts (via *retweeting*) and *then* posting pro-ISIS content themselves, we derived the $\Delta(a_{h1} - a_{h2})$-distribution using all users that fall within the *intersection* of the H1 and H2 users' sets. For each user in this intersection set ($u \in U_{H1} \cap U_{H2}$) we measured the difference (in days) between their H2 activation point (a_{h2}) - i.e. when they first post pro-ISIS rhetoric themselves - and their H1 activation point (a_{h1}) - i.e. when they first shared content from pro-ISIS accounts. Figure 2(c) presents the distribution of $\Delta(a_{h2} - a_{h1})$. We note that this distribution has a right skew indicating that the majority of users post pro-ISIS terms before then going on to *share* content from pro-ISIS accounts - note that we only have 64 users within intersection of H1 and H2 users.

Detecting Behaviour Divergence

Having detected the activation points of users within both the H1 and H2 hypotheses' sets, we then moved on to examine what happens once users have become *activated*: **RQ2**: *What happens to Twitter users before they exhibit radicalised behaviour, and also after such exhibition?* As behaviour is a fairly abstract concept, we operationalise its measurement through three *dimensions*: (i) the lexical terms used by a user (i.e. non-stop word terms published in his/her tweets), (ii) the users whose content the user has *shared* (i.e. propagated through his network), and (iii) the users that the user has *mentioned*. Each dimension, which we refer to as *lexical*, *sharing*, and *interactions* respectively, in essence forms a discrete probability distribution that we can derive from a given half-closed time interval (i.e. $[t, t') : t < t'$). Each distribution is then derived from the relative frequency distribution of the user's behaviour within the allotted time window: for instance, the lexical dimension's distribution ($P^{L}_{[t,t')}$) is the relative frequency distribution of terms used within the user's tweets within the time window.[5] As we are dealing with both Arabic and English tweets, we ran a process of *transliteration* on the former to convert Arabic script to English unicode characters, thereby allowing for both languages to be handled using the same *base* language.

In order to examine whether a user's behaviour has changed once activated we computed the *relative entropy* (aka. Kullback-Leibler/KL divergence) over three time windows. Each time window has a *midpoint* (m), this midpoint then forms the boundary from which a given behaviour dimension has two probability distributions computed (one before the midpoint, and one after the midpoint). Let $P_{[t,m)}$ denote the distribution prior to m, and $Q_{[m,t')}$ denote the distribution on and after m, then the relative entropy is computed using P and Q as follows:

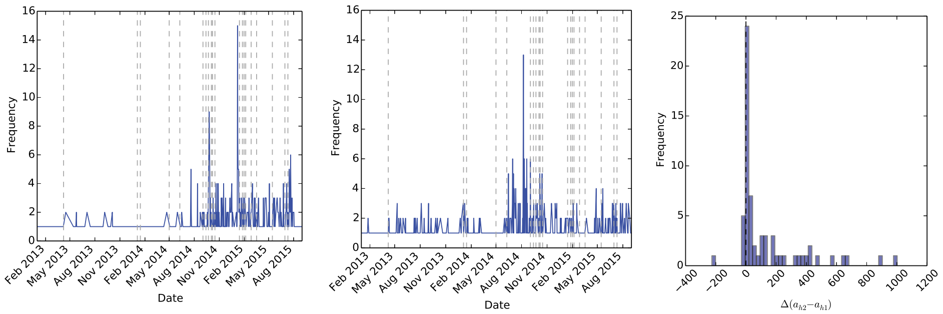
$$H(Q||P) = \sum_{i} P(i) \log \frac{P(i)}{Q(i)} \tag{1}$$

As mentioned above, we measured the relative entropy over three windows, these were as follows:

1. *Activation Window*: the midpoint (m) of the window is the given user's activation point (i.e. a_{h1} or a_{h2}), and we set the bounds of the window by going back k days from m.
2. *Pre-Control Window*: the midpoint of the window is $2k$ days *back* from the activation point of the user, and the bounds are set to $[a - 3k, a - k)$.
3. *Post-Control Window*: the midpoint of the window is $2k$ days *forward* from the activation point of the user, and the bounds of the window are set to $[a + k, a + 3k)$.

Hence, our experimental setting provides three *non-overlapping* time windows over which we could compute the relative entropy of user behaviour (lexical, sharing, interactions). For users labelled as pro-ISIS by H1 and H2 we computed their three relative entropy values over the three

[5] The *sharing* and *interactions* distributions are computed in the same manner, using the relative frequencies of users whose content is shared and users mentioned respectively.



(a) H1 Users

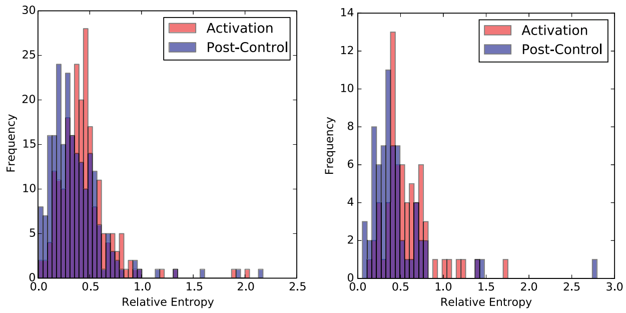
(b) H2 Users

(c) Delta distribution

Figure 2: Frequency of activations for H1 users (Figure 2(a)) and H2 users (Figure 2(b)) with key ISIS/ISIL events marked, and the delta distribution measured between H1 and H2 activation points (in days) over the intersection of H1 and H2 users (Figure 2(c)).

respective time-windows and then tested the null hypothesis that the distribution of relative entropy values did not differ between consecutive windows - we set $k = \{25, 50, 100\}$ - using the paired-sign test.[6]

Table 3 contains the significance probabilities obtained from applying the paired signed test to consecutive windows' relative entropy distributions across users labelled using H1 and H2 and the three behaviour dimensions. The results indicate that users *significantly* diverge from the *activation* window into the *post-control window* based on the language they use. In fact, when we examine the distribution of relative entropies in the activation window versus the post-control window (showing for only $k = 25$ days, Figure 3), we find that users exhibit a large *divergence* in their language once *activated* - within the activation window - whereas the post-control window's relative entropies are lower. This suggests that the activation process of users results in a clear *shift* in the behaviour, in fact the values in parentheses in Table 3 indicate that the activation window yields (in general) the largest relative entropies and thus the greatest change in behaviour through the window.



(a) H1 Users - $k = 25$

(b) H2 Users - $k = 25$

Figure 3: Distribution of lexical behaviour dimension's relative entropy distributions within the activation window and within the post-control window. The former's distribution has a higher *location* than the latter, indicating that the activation window contains a greater *change* in behaviour.

To look at the language that pro-ISIS users adopt before they are activated, within their activation window, and after they are activated, we induced bag of words models over the pre-control, activation, and post-control windows respectively for each user. We then combined these bags of words to induce term-frequency vectors for all pro-ISIS users who become activated at some point. We also computed the average sentiment of each term as follows: for each Tweet, we computed its sentiment by matching the Tweet's words against the MPQA lexicon for English (Wilson, Wiebe, and Hoffmann 2005) and the ArSenL lexicon for Arabic tweets (Badaro et al. 2014), and derived the overall sentiment from the average sum of the opinionated terms within it - this then set the sentiment of each term within the Tweet (i.e. positive or negative, and the degree of polarity). Each term's average sentiment - within the bag of words models merged together across the users - was then computed. Looking at the terms used by users before being activated (Figure 4(a)) shows that the majority of topics users discuss focus on politics, where words like Syria, Israel and Egypt are mentioned in a negative context and with high frequency. Once users become activated and thereafter (Figures 4(b) and 4(c)) it is clear that religious words (e.g. Allah, muslims, quran) become more popular. We also note that here, the term *ISIS* is mentioned in a negative context (i.e. the red colouring indicates a negative sentiment), this is likely due to pro-ISIS users not referring to Islamic State using the abbreviation 'ISIS' - hence it is likely that such usage is derogatory towards those using the term.

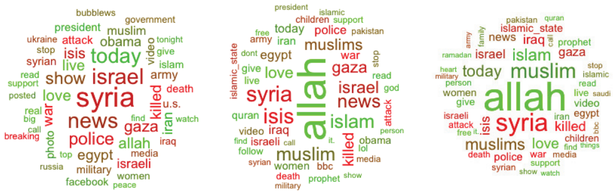
Pathways to Activation

Our exploratory analysis of users' behaviour prior to, during, and post-activation has revealed some interesting insights into the divergent properties of user behaviour over those periods. In this section we now move onto examining

[6] N.b. As our paired-samples data is neither normal nor symmetric, we could not use the Wilcoxon signed-rank test nor the paired two-sample T-test

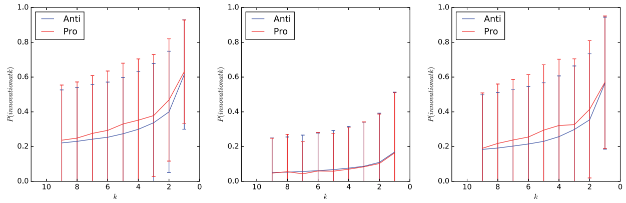
Table 3: Sign test significance probabilities produced when testing the null hypothesis that users' consecutive windows' relative entropies are equal (i.e. that behaviour remains *stable*) across the three behaviour dimensions: lexical terms used (Lexical), users' content that is shared (Sharing), and users interacted with (Interactions). Up arrows in parantheses indicate which window's distribution had the greater location.

		H1		H2		H1 ∪ H2	
	k	$H(A\|\|Pr)$	$H(Po\|\|A)$	$H(A\|\|Pr)$	$H(Po\|\|A)$	$H(A\|\|Pr)$	$H(Po\|\|A)$
Lexical	25	$0.493(\uparrow Pr)$	$0.000(\uparrow A)$	$0.025(\uparrow A)$	$0.000(\uparrow A)$	$0.503(\uparrow A)$	$0.000(\uparrow A)$
	50	$1.000(\uparrow A)$	$0.000(\uparrow A)$	$0.243(\uparrow A)$	$0.040(\uparrow A)$	$0.472(\uparrow A)$	$0.000(\uparrow A)$
	100	$0.832(\uparrow Pr)$	$0.000(\uparrow A)$	$0.004(\uparrow A)$	$0.004(\uparrow A)$	$0.399(\uparrow A)$	$0.000(\uparrow A)$
Sharing	25	$1.000(\uparrow A)$	$0.508(\uparrow A)$	$1.000(\uparrow A)$	$0.012(\uparrow A)$	$0.824(\uparrow A)$	$0.012(\uparrow A)$
	50	$0.453(\uparrow A)$	$1.000(\uparrow A)$	$0.219(\uparrow A)$	$1.000(\uparrow A)$	$0.022(\uparrow A)$	$1.000(\uparrow A)$
	100	$0.375(\uparrow Pr)$	$1.000(\uparrow A)$	$1.000(\uparrow Pr)$	$1.000(\uparrow A)$	$0.549(\uparrow Pr)$	$1.000(\uparrow A)$
Interactions	25	$0.181(\uparrow Pr)$	$0.040(\uparrow A)$	$0.648(\uparrow A)$	$0.001(\uparrow A)$	$0.614(\uparrow Pr)$	$0.001(\uparrow A)$
	50	$0.092(\uparrow Pr)$	$0.276(\uparrow A)$	$1.000(\uparrow A)$	$1.000(\uparrow A)$	$0.227(\uparrow Pr)$	$0.403(\uparrow A)$
	100	$0.262(\uparrow Pr)$	$0.011(\uparrow A)$	$0.006(\uparrow A)$	$0.039(\uparrow A)$	$0.798(\uparrow Pr)$	$0.002(\uparrow A)$



(a) Pre-Control (b) Activation (c) Post-Control

Figure 4: Word clouds of the top-50 terms published by pro-ISIS users before, during and after their activation, the colour indicates the sentiment attached to the term - with red being negative, and green being positive. We note that political topics make way to religious topics.



(a) Lexical (b) Sharing (c) Interactions

Figure 5: Relative frequencies of innovations (Figures 5(a), 5(b), and 5(c)) of users from the anti-ISIS and pro-ISIS groups (union of H1 and H2 users) in the k weeks prior to their *activation* with either a pro or anti-ISIS position. Both pro and anti-ISIS users exhibit similar trends.

in more detail what becomes of users prior to their *activation*.

Behaviour Prior to Activation

Understanding what users go through prior to their activation, either with the adoption pro-ISIS rhetoric or the sharing of content from known pro-ISIS accounts, could reveal how users' behaviour changes over time. To follow this line of thinking, we computed the relative entropies across the three behaviour dimensions of *both* pro and anti-ISIS users in the k weeks *prior* to their activation point.[7] We found (plots omitted for brevity) how there was little to discern between the user groups in terms of behaviour divergence on a week-by-week basis. However, after calculating the weekly innovation relative frequency (i.e. the proportion of unique terms, users retweeted, and users interacted with) for those k weeks prior to activation (Figures 5(a), 5(b), and 5(c)) it became apparent that users become more *amenable* to innovation the closer they get to their point of activation - this is also apparent across the two groups (pro and anti-ISIS).

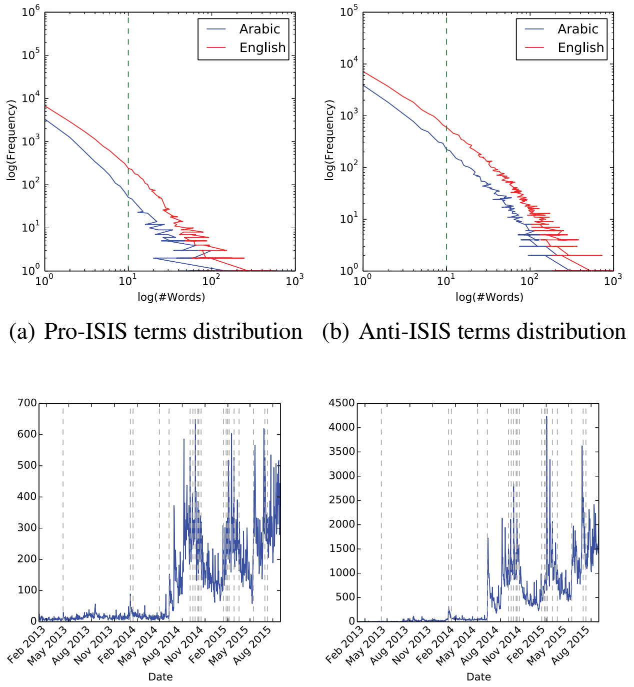
Focussing more now on the terms that users' uses that signify both pro and anti-ISIS content, we see that anti-ISIS terms are more commonly used than pro-ISIS terms (Figures 6(a) and 6(b)), however their emergence over time follows similar trends (Figures 6(c) and 6(d)) where large spikes are evident around the period where the majority of users' becoming activated. The large surge in update of both pro-ISIS terms and the dramatic increase in the innovation relation frequencies in the k weeks prior to activation, as $k \rightarrow 0$, suggests that users become more susceptible to *adoption* the closer to being activated they are.

Influences on Pro-ISIS Term Adoption

Having found that users, prior to their *activation*, exhibit a relative increase in innovative behaviour we then sought to examine what *influences* users to adopt specific innovations: namely, pro-ISIS terms (**RQ3**: *What influences users to adopt pro-ISIS language?*) Our aim here was to *disentangle* different *influence* factors that govern the adoption process. For instance, if a user adopts a known pro-ISIS term and then goes on to use that term several times (and thus becomes activated) then we are interested in knowing *what* influenced the user after he/she is *exposed* to the term to then begin using it.

To follow this avenue of investigation, we implemented and applied an adaptation of Goyal et al.'s (Goyal, Bonchi, and Lakshmanan 2010) *General Threshold Diffusion Model*.

[7] To identify anti-ISIS signals, we used the same process as pro-ISIS signals when applying H2 - i.e. more anti than pro-ISIS terms, and minimum of 5 uses of a single anti-ISIS term.



(a) Pro-ISIS terms distribution (b) Anti-ISIS terms distribution

(c) Pro-ISIS terms + events over time (d) Anti-ISIS terms + events over time

Figure 6: The log-log frequency distributions of Pro (Figure 6(a)) and Anti-ISIS (Figure 6(b)) terms for both Arabic and English - where pro-ISIS term usage is markedly less than anti-ISIS term usage, and the use of such terms over time (Figure 6(c) and Figure 6(d)) showing dramatic increase from 2014 onwards.

Our goal here was to compute, for a given pro-ISIS term, the adoption probability p_u that a given user u will adopt the term for the first time and go on to use it (i.e. become activated). We assume that the diffusion process unfolds over time such that when a user v adopts the term he/she then has a probability of *influencing* all other non-activated users into adopting the term: hence, we can decompose this as a problem of calculating the influence probability of v on u as $p_{v,u}$. Calculation of the probability that u adopts the term is as follows: [8]

$$p_u(A) = 1 - \prod_{v \in A} \left(1 - p_{v,u}\right) \tag{2}$$

Where A is the set of all activated users that have adopted the term prior to t_u (the time at which u adopts the term). This neat formulation ensures that $p_{v,u}$ can be varied across different *modalities* thereby quantifying the influence that v has upon u using different means. As stated above, we are interested in exploring how different influence factors impact a user's adoption probability, therefore we created three influence probability functions each capturing a different modality of influence:

1. **Lexical homophily**: Given users u and v and activation point of v defined by t_v, we generated a bag of words model, after filtering-out stopwords, for both users using only their Tweets prior to t_v. From these models, we then computed frequency vectors $\mathbf{x}_u$ and $\mathbf{x}_v$, and calculated the angular cosine similarity between the vectors.[9] Inclusion of this factor was to examine whether the similarity in language used between users has an influence on them adopting a pro-ISIS term.
2. **Sharing homophily**: Following the same process as lexical homophily, we derived the vectors for u and v by recording the frequencies of *sharing* from unique Twitter accounts prior to t_v. And as above, we then calculated the angular cosine similarity between the vectors. Inclusion of sharing homophily was intended to discern whether the diffusion of information by u and v from the same accounts impact u adopted a pro-ISIS term, this is somewhat pertinent given the *endorsement*-like effect of *retweeting*.
3. **Interactions homophily**: Here we calculated user vectors based on the number of interactions (not sharing) with previous users by u and v prior to t_v, and then calculated the angular cosine similarity between these vectors. Our rationale behind exploring this influence factor was to assess the impact that common communications may have on term adoption.

The adoption probability (p_u) of a given user (u) is calculated based on the pseudocode in Algorithm 1. We begin the process (line 1) by gathering the set of users that have been activated with a given pro-ISIS term (U_w) and generate a balanced set of users to generate the adoption probabilities for - which we define as U_p and is our *to-process* user set. This *balanced* set contains a 50:50% split of users who adopt w and those users who do not adopt w, we generate these latter users by randomly sampling $|U_w|$ users from our collected dataset who are not found to be exhibit pro-ISIS signals (based on H1 or H2) or anti-ISIS signals (i.e. have used anti-ISIS rhetoric). We then gathered the user-time tuples from tweets citing w (line 2) and primed the set of activated users (A) and the result set (R) to be empty sets. Our model then runs through the time-ordered user-time tuples and checks if the user (v) has been activated before, if not then we compute the influence probability between v and each user who is yet to be activated ($\forall u \in U_p$) and update the adoption probability of u using the update rule on line 10 - given that the adoption probability (Equation 2) is both *submodular* and *monotonic*. For v, after calculating the influence probabilities between v and all other users that are to be processed, the results table is then updated (lines 14-17) with the probability of v adopting w and the outcome: 2 if v is the first adopter of w and 1 otherwise.

[8] Note that a property of this joint probability function is that it is *submodular* and *monotonic* - this ensures that including a new probability ($p_{v,u}$) will only increase p_u. We exploit this property to enable parallel processing as we document below.

[9] We use the angular similarity that constrains the co-domain to the closed interval $[0, 1]$.

Algorithm 1 Calculation of per-user adoption probabilities. **Input**: Pro-ISIS term w, users activated with term U_w. **Output**: Set of result tuples $(u, p_u, adopted) \in R$.

```
1:  $U_p \leftarrow$ balanced set of users to process
2:  $T \leftarrow$ tweet tuples citing $w$
3:  $A, R \leftarrow \emptyset$
4:  for $(v, t_v) \in sorted(T)$ do
5:      if $v \notin A$ then
6:          $A \leftarrow A \cup v$
7:          $U_p \leftarrow U_p \setminus v$
8:          for $u \in U_p$ do
9:              if $(u, p_u, .) \in R$ then
10:                 $p_u \leftarrow p_u + (1 - p_u) * p_{v,u}$
11:                 Update $R$ with $(u, p_u, 0)$
12:             else
13:                 $R \leftarrow R \cup (u, 0, 0)$
14:     if $R == \emptyset$ then
15:         $R \leftarrow R \cup (v, 0, 2)$
16:     else
17:         Update $R$ with $(v, p_v, 1)$
18: Return $R$
```

Parallel Processing The sequential nature of calculating adoption probabilities using algorithm 1 can result in lengthy computation times. To speed up the process, the algorithm was parallelised by implementing the code in Java and using Apache Spark to divide processing across a 12-machine cluster (with 30 CPUs and 270Gb RAM) as follows: we first split the processing up by parallelising at the innovation level - so one task runs per innovation. Second, we created a multi-threaded version of lines 8-13 so that pairwise influence probabilities between v and each user from U_p were computed in parallel. All data was loaded into HDFS and HBase tables to ensure quick lookup - in particular, HBase tables containing users' posts were indexed on user identifiers to enable quick concurrent access.

Experiments Our aim here was to compare how social homophily, sharing homophily, and interaction homophily fare when calculating adoption probabilities. To enable such a comparison, for all pro-ISIS terms (both Arabic and English) we calculated users' adoption probabilities resulting in the derivation of a set of result tuples $(u, p_u, adopted) \in R$ for each term and influence factor. Using these tuples we then judged the accuracy of the different influence factors, which in turn dictate the value returned for $p_{v,u}$, by calculating the area under the Receiver Operator Characteristic curve (ROC) - which measures accuracy based on true positive and false positive rates. Area under the curve was calculated for each model by deriving (FPR, TPR) pairs as the confidence of the given model was increased through the interval $[0, 1]$ at steps of 0.05, plotting the resultant curve, and then calculating the area under it - where a value of 0 indicates poor performance, 1 is perfect, and 0.5 is equivalent to randomly guessing.

Our results are shown in Table 4 for individual pro-ISIS terms and together with the macro and micro-averages of the models.[10] It is clear that when calculating influence between users based on the sharing homophily that we achieve the best performance, followed by interactions homophily; hence, the *social homophily* has a stronger bearing on influencing pro-ISIS term adoption than merely lexical similarity. This suggests that sub-communities within Twitter, from which content arises and is passed on (via *sharing*) and with whom those individuals are interacted, exist that both *adopters* and *future adopters* are connected and are listening to. We also note that our models function better for Arabic terms than English terms.

Discussion

In this paper we investigated how users develop to exhibit signals of pro-ISIS (radicalisation) behaviour. Our findings have implications for researchers spanning digital humanities, religious studies and computational social science interested in examining the development of users prior to adopting a *potentially* radicalised stance. The computational techniques used within our work are also of interest to researchers working on modelling diffusion over large-scale time-series data. We now reflect on what our research has revealed, the limitations of our study, and how we can proceed this work forward in the future. One core limitation, or rather result, of our work is the sparsity of users that we identify as exhibiting *signals of radicalisation* with only 508 users identified based on their sharing of content from known pro-ISIS accounts (H1), 208 users based on pro-ISIS language (H2), and 64 users in the intersection of the sets of H1 and H2 users. Given that we focus exclusively on 154K users from European countries these numbers do corroborate with those from Berger and Morgan (Berger and Morgan 2015) - where the authors found on average $<$150 pro-ISIS users per Western European country in their sample. However, this sparsity in the sample does lead one to question the *generalisation* capability of our work, hence: our future work intends to repeat the process described within this paper in order to triangulate our findings via a repeated study.

Unlike prior work by Berger and Morgan (Berger and Morgan 2015) we did not use any information contained within tweets' hyperlinks, this is despite Berger and Morgan dereferencing such links and finding links to ISIS propaganda. Our future work will rectify this by adding an additional hypothesis for identifying signals of radicalisation based on users acting as *sources* for such content - i.e. when acting as *initial adopters* of the URLs. Dereferencing of hyperlinks would also allow one to investigate the link between the timing of salient ISIS events involving Western countries (see Table 2) and users adopting language associated with shared material; we could also enable the diffusion model to account for the *recency* of key external events and their impact on pro-ISIS term adoption.

Throughout this research we have adopted an exploratory data mining approach by collecting data and then analysing

[10] The macro-average is determined by computing per-term ROC values and then averaging, while the micro-average is computed by merging all result tables together and determining the ROC value from the merged table.

Table 4: Performance of different influence probability modalities when predicting pro-ISIS term adoption probabilities. Macro-ROC includes is per-term ROC values averaged, while Micro-ROC is computed using all $(prob, outcome)$ tuples.

Term	Description	Lexical	Sharing	Interactions
الَامَارة الَاسلَامية	The Islamic State	0.500	0.611	0.611
الدولة الَاسلَامية	The Islamic State	0.479	0.673	0.579
اْلْدْوْلْةْ اْلَاءْسْلَامْيْةْ	The Islamic State	0.469	0.660	0.568
Apostate	Person who criticises/rejects Islam	0.387	0.485	0.484
Khilafah	The Islamic State	0.349	0.482	0.519
Shirk	Blasphemy	0.451	0.481	0.487
Ummah	Denoting '*nation*'	0.412	0.471	0.493
Macro-ROC		0.433(±0.054)	0.551(±0.084)	0.535 ± 0.047)
Micro-ROC		0.476	0.602	0.551

it based on our hypothesised signals of radicalisation, we do not attempt to associate online behaviour (i.e. sympathising with the ISIS cause) and offline actions. That said, our investigation into the adoption of pro-ISIS terms as a diffusion process has revealed some interesting trends that cyber-security and intelligence services might take note of. Firstly, we found that social dynamics play a *strong* role in term uptake where users are more likely to adopt pro-ISIS language from users with whom they share many interacted users (either via having communicated with those users beforehand, or shared content from them). This finding suggests that such common users act as *bridges* between the term adopter and the future adopter, and could thus warrant further inspection. Secondly, prior to being *activated* users go through a period of significant increase in adopting innovations (i.e. communicating with new users and adopting new terms), this clear increase suggests that users are *rejecting* their prior behaviour and escalating this further until becoming activated - in a similar manner to that explained in (King and Taylor 2011).

Acknowledgment

This work was supported by the EU-FP7 project SENSE4US (grant no. 611242).

References

Badaro, G.; Baly, R.; Hajj, H.; Habash, N.; and El-Hajj, W. 2014. A large scale arabic sentiment lexicon for arabic opinion mining. *ANLP 2014* 165.

Bartlett, J., and Miller, C. 2012. The edge of violence: Towards telling the difference between violent and non-violent radicalization. *Terrorism and Political Violence* 24(1):1–21.

Bazan, S.; Saad, S.; and Chamoun, M. 2015. Infowar on the web: When the caliphate goes online. In *ACM Web Science Conference*.

Berger, J., and Morgan, J. 2015. The isis twitter census: Defining and describing the population of isis supporters on twitter. *The Brookings Project on US Relations with the Islamic World* 3:20.

Berger, J. M. 2015. Tailored online interventions: The islamic state's recruitment strategy. *Combatting Terrorism Center*.

Bermingham, A.; Conway, M.; McInerney, L.; O'Hare, N.; and Smeaton, A. F. 2009. Combining social network analysis and sentiment analysis to explore the potential for online radicalisation. In *Social Network Analysis and Mining, 2009. ASONAM'09. International Conference on Advances in*, 231–236. IEEE.

Edwards, C., and Gribbon, L. 2013. Pathways to violent extremism in the digital era. *The RUSI Journal* 158(5):40–47.

Fleiss, J. L.; Levin, B.; and Paik, M. C. 2013. *Statistical methods for rates and proportions*. John Wiley & Sons.

Goyal, A.; Bonchi, F.; and Lakshmanan, L. V. 2010. Learning influence probabilities in social networks. In *Proceedings of the third ACM international conference on Web search and data mining*, 241–250. ACM.

Hall, J. 2015. Canadian foreign fighters and isis.

King, M., and Taylor, D. M. 2011. The radicalization of homegrown jihadists: A review of theoretical models and social psychological evidence. *Terrorism and Political Violence* 23(4):602–622.

Klausen, J. 2015. Tweeting the jihad: Social media networks of western foreign fighters in syria and iraq. *Studies in Conflict & Terrorism* 38(1):1–22.

O'Callaghan, D.; Prucha, N.; Greene, D.; Conway, M.; Carthy, J.; and Cunningham, P. 2014. Online Social Media in the Syria Conflict: Encompassing the Extremes and the In-Betweens. In *Proc. International Conference on Advances in Social Networks Analysis and Mining (ASONAM'14)*.

Torok, R. 2013. Developing an explanatory model for the process of online radicalisation and terrorism. *Security Informatics* 2(1):1–10.

Wilson, T.; Wiebe, J.; and Hoffmann, P. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, 347–354. Association for Computational Linguistics.

Winter, C. 2015. Documenting the virtual 'caliphate'. *Quilliam Foundation*.