Tuning in to Terrorist Signals

Shaun Wright

Thesis submitted for the degree of Doctor of Philosophy

Royal Holloway

University of London

2017

Declaration of Authorship

I, Shaun Philip Wright, hereby declare that this thesis and the work presented in it is entirely my own—with the exception of the partly co-authored work clearly outlined below. Where I have consulted the work of others, this is always clearly stated.

With the exception of *Chapter 4. How humans transmit language*, where my contribution is outlined below, the work presented in this thesis is my own. In Chapters 5, 6 and 7, I collected all of the data, performed all of the analyses, wrote the first draft of each manuscript and led the writing of the manuscripts. All supervisors (David Denney (DD), Alasdair Pinkerton (AP) , Vincent AA Jansen (VAAJ), Peter Adey (PA) and John Bryden (JB)), however, were involved in the conception of the work and DD, AP, VAAJ and JB edited and critiqued the manuscripts.

Chapter 4. How humans transmit language was co-authored with JB and VAAJ. The data for this chapter was downloaded by JB in 2009. The initial draft of the work was carried out by me. That initial analysis now constitutes the results in Fig. 4. and the initial literature review informs both the current draft and the language section in the literature review of this thesis. After subsequent discussion between JB, VAAJ and myself, *VAAJ devised the model in the supplementary material* of Chapter 4 and *JB carried out the modelling analysis (Fig. 3.)*. The analysis forming Fig. 2. was initially conceived and carried out by me on a sample of around 100 words and *subsequently repeated by JB on the 1,000 words in the manuscript*. Although I wrote the first draft and reviewed the literature, *JB re-started, formatted, produced figures for and led the writing of, the current draft*. All authors, including myself, edited and critiqued the manuscript. Overall, all authors, including myself, contributed equally to this chapter.

Signed _____

Dated _____

Abstract

Twitter, social media and big data promise much in terms of terrorist signals amenable to analysis. As, however, these signals are noisy, subjectively ambiguous and new, this thesis addresses four questions that are key to reliably 'tuning in' to these signals. Each chapter uses big data to investigate patterns too subtle to have been amenable to prior study, with the importance of controlling for the noise associated with big data a central theme running through the thesis.

Chapter 1 introduces the work, Chapter 2 reviews the relevant literature and Chapter 3 introduces and discusses the overarching methodology.

Chapter 4 considers the validity of inferring information about users from their Twitter language and tweets. I demonstrate that language can be horizontally transmitted and inherited; with behaviour and interactions leading to and predicting, changes in language. This extends previous work with small sample work that did not exclude imitation.

In Chapter 5, I characterise jihadist-linked accounts that resurge back from suspension —as identified with novel methods. I show that suspension is less disruptive than previous case studies implied, but that pseudoreplication has been underestimated (Wright, 2016).

Having demonstrated the scale of resurgence, Chapter 6 tests whether automated machine methods can improve identification. I develop a text similarity based model and validate it against human-annotated data.

The final research chapter, Chapter 7, tackles noise in big data when inferring information about events in the offline world. Extending similar work, I evaluate computational and human coded predictions of how positive geopolitical events are for Daesh. I demonstrate that while the Baqiya family tweets differently on different types of day, most patterns emerge as easily by chance in the negative control data.

The work is novel as although some attempts have been made to address the questions in this thesis—or similar ones—using case studies, small samples and laboratory studies, all of these suffer limitations. Some studies have not asked the exact same question, some conclusions have been insufficiently supported with evidence and others have simply been beyond the reach of existing methods.

Together, the pieces of work in this thesis shows that computational analysis of big data enables tuning in to subtle signals and sometimes reveals conclusions that contradict less developed research. Control noise, however, often contains as many patterns and thus, future studies should pay particular attention to their methodologies when using noisy, subjective, social media data.

Acknowledgements

Acknowledgements must first go to Royal Holloway University of London itself. Funding for my fees and studentship came from Royal Holloway's three year Reid Scholarship 2013/14, which was designed with the fantastic aim of promoting interdisciplinary work between departments. Not only is such collaboration lauded as the future of academia, but was perfectly designed for me and my preference for moving from one related topic to another. I am very grateful for the opportunity. The support from Klaus Dodds, research champion for the Security and Sustainability research theme, has also been invaluable.

The next most important acknowledgement is for the contribution of my supervisors—Vincent Jansen, David Denney, John Bryden, Alasdair Pinkerton and Peter Adey. For three years, Vincent has been just a desk away and has been the most patient and helpful supervisor one could hope for—my deepest thanks to him. My thanks also to John, for sharing his technical, computational and London expertise and for his previous work, upon which this work builds; to David, who not only introduced me to the social sciences, but gave me the immensely enjoyable privilege of teaching alongside him—as well as generously giving me a lift home on multiple occasions; to Al, whose door was always open and offered valuable advice long before we were finally able to formally add him to the supervisory team; and to Pete, who, although he had a busy few years, provided fresh criticisms at each update meeting. I am very grateful to you all.

I would also like to thank various members of the School of Biological Sciences: my advisor Francisco, postgraduate research administrator Tracey and all the members of the EEB (Ecology, Evolution and Behaviour) research theme for their feedback at the weekly research seminars.

Finally, the support from Nicci, my family and the new friends I made at Royal Holloway—especially Lauren, Kate and Rosie—has not only kept me sane, but means that I will look back on these three years as some of the best in my life.

"Live long and die happy."

Contents

1. Introduction

Scientific, computational and social media data methods have not been widely applied to tackle social and geopolitical questions. In part, this is because subjective social phenomena are difficult to study with machine methods. Large volumes of representative data from online social networks have, however, made previously hard to study social phenomena amenable to investigation. Nonetheless, this data is still subjective and noisy, presenting difficulties with the application of standard machine methods.

This thesis, therefore, is about an attempt to rigorously study social phenomena using significant volumes of online social media data. I apply computational and machine methods to novel questions, adapting them to develop solution to overcome some of the ambiguity, subjectivity and noise problems. In particular, the questions relate to understanding terrorism and political violence, data for which have become commonplace on social media. From a social scientific point of view, these are existing problems and areas of study, but the machine methods are novel.

In particular, the approach used to interface between social phenomena and machine methods is through developing tools to study language patterns. Inferring information about users from their Twitter language and tweets is possible as language both conveys information about the person producing it and changes throughout a person's life as they inherit language through communication. The first research chapter—chapter 4—considers the validity of this approach, demonstrating that language can be horizontally transmitted and inherited; with behaviour and interactions leading to and predicting, changes in language. This extends previous work with small sample work that did not exclude imitation (Branigan *et al.*, 2011; Brennan, 1996; Christopherson, 2011; Danescu-Niculescu-Mizil *et al.*, 2011; De Looze *et al.*, 2011; 2014; Hemphill and Otterbacher, 2012; Steinhauser *et al.*, 2011).

After supporting the use of language patterns as a proxy for social patterns, the following research chapter—Chapter 5—characterises jihadist-linked accounts that resurge back from suspension, as identified with novel methods. I show that suspending

users is less disruptive than previous case studies have implied (Stern and Berger, 2015; Berger and Perez, 2016), but that biases such as pseudoreplication have been underestimated (Wright *et al.*, 2016).

Having demonstrated the scale of the problem caused by resurgence, chapter six investigates whether automated machine methods can improve the identification of the multiple accounts created by the same person (Berger and Morgan, 2015; Chatfield *et al.*, 2015; Magdy *et al.*, 2015). I develop a text similarity based model and validate it against human, crowd sourced performance.

The final chapter tackles noise in ambiguous big data when inferring information about subjective events in the offline world. Extending work by Magdy *et al.* (2015), I evaluate computational and discourse analysis predictions of the parity of geopolitical events relating to Daesh. I demonstrate that whilst the Baqiya family tweets differently on different types of day, most patterns emerge as easily by chance in the negative control data.

The work is novel as although some attempts have been made to address the questions in this thesis—or similar ones—using case studies (Stern and Berger, 2015; Berger and Perez, 2016), small samples (Christopherson, 2011; De Looze *et al.*, 2011; 2014) and laboratory studies (Branigan *et al.*, 2011; Brennan, 1996; Steinhauser *et al.*, 2011), all of these suffer limitations. Some studies have not asked the exact same question (Magdy *et al.*, 2015), some conclusions have been insufficiently supported with evidence (Stern and Berger, 2015; Berger and Perez, 2016) and others have simply been beyond the reach of existing methods (Berger and Morgan, 2015; Chatfield *et al.*, 2015; Magdy *et al.*, 2015).

The overarching conclusion of this thesis is that this approach works—machine and other quantitative, computational methods have a lot to offer to the social sciences, but caution must be exercised as, without rigourous experimental design, false positives easily emerge.

2. Literature Review

2.1 Overview

This literature review evaluates the two distinct bodies of literature upon which the work in this thesis is built. This thesis aims to show that analysis of the language contained in large volumes of social media data enables us to tune in to subtle terrorist signals on Twitter, thereby revealing novel conclusions about terrorism.

Firstly, therefore, historical context and literature from the field of terrorism studies is important. As this thesis addresses sociological, scientific and machine-method problems framed in the way that they are most relevant to scholars of terrorism and uses data sampled from accounts associated with extremists on Twitter, an overview of the terrorism landscape is important. I start with a review of the definitions of what terrorism is and who becomes a terrorist. I then give an outline of the historical context, offline and on Twitter, at the point when this work was conceived—2013—and then cover the fast-changing trends that have been documented over the duration of the work on this thesis.

The second area of background literature informing this thesis is language research. Language is important to this thesis as it provides the model system and proxy to uncover novel patterns used in each study. The literature review considers both the literature on language as a proxy for associated sociological and behavioural characteristics and our biological and psychological / linguistical understanding of how communication can lead to language change and evolution.

2.2 Terrorism

Terrorism is one of the greatest threats that our world today faces. According to the MI5 website: *"terrorism is the biggest national security threat that the UK currently faces"* and in a November 2014 interview, the then UK Home Secretary, Theresa May, said that the threat was *"greater than it has been at any time before or after 9/11"* (Dominiczak, 2014).

Those statements proved prescient as in January 2015 there were three days of attacks in Paris, beginning at the offices of satirical magazine Charlie Hebdo (BBC, 2015). Also that month, the UK police terror alert was raised to its highest ever level (Dodd *et*

al., 2015) and the Director General of MI5 described *"complex and ambitious plots [aiming to] cause large scale loss of life"* (ITV Report, 2015a), with then UK Prime Minister David Cameron stating that the terrorist threat was *"the greatest concern that I have as Prime Minister"* (ITV Report, 2015b).

To illustrate the threat from terrorism, for just the first two years of this doctoral work, 2014 and 2015, the Global Terrorism Database (National Consortium for the Study of Terrorism and Responses to Terrorism (START), 2016) documents 21,646 incidents of domestic or international terrorism. Although Daesh (also known as ISIS, ISIL, IS and the Islamic State, but, for reasons outlined in the introduction to Daesh in section 2.2.4, henceforth referred to as Daesh) dominates our headlines, it is recorded as the perpetrator in only 2,494 (11.5%) of those attacks and of those, only 9 (0.4%) occurred outside of the Middle East & North Africa region.

2.2.1 Defining terrorism

There is no universally accepted definition of terrorism (Clarke, 2009; Schmid, 2011). Given the age and public notoriety of this phenomenon, this might seem strange. Terrorism has been a very public phenomenon; from the knife attacks by the highly organised, religious sect the *Sicarii* during crowded festivals in Palestine from 66-73 C.E. (Laqueur, 2012), through *la Terreur* in 18th Century revolutionary France, bombings by the Zionist Irgun (Walton, 2014), the Palestinian Liberation Organisation (PLO) and the Irish Republican Army (IRA) in the first, second and second halves of the 20th Century respectively, to the rise of Islamist terrorism dominated first by al-Qaeda and now Daesh. Publicity is, after all, often a principal goal of the attackers. Terrorism therefore features frequently in the media, alongside public discussion by 'experts' and consequently, most people would be fairly confident in recognising, labelling and condemning a hypothetically or historically presented act as terrorism if they considered it to be so (Meisels, 2008). At first glance, terrorism appears to be an intuitively recognisable act, but it is very challenging to define.

Why is having a definition of terrorism so important? Firstly, the fact that there is disagreement over what constitutes terrorism makes it vital that any book, article or academic research investigating or evaluating terrorism begins with a clear statement of

the boundaries under consideration. The definition adopted may well be contentious, but it is nonetheless necessary to allow others, within the scope of the definition given, to both review their work fairly and correctly *and* integrate it accurately with other research. Secondly, beyond academia, legal frameworks require a definition, for example for the enforcement of international agreements. Just one problem caused by the lack of a universally accepted definition is the extradition of terrorists (Ganor, 2010). Although most nations have agreed to extradite terrorists to the country in which they face charges, the lack of an international definition of who is a terrorist provides a loophole (either for honest disagreement or more deliberate abuse) for countries to avoid this obligation (Ganor, 2010). New and existing legal, operational and academic frameworks all require terrorism to be defined (Clarke, 2009) and whilst some cases might simply need precise articulation of *'a'* definition, others require mutual acceptance of a single definition. In academic studies, including this thesis, it is often the former.

As the next few paragraphs will set out, opposition to definitions of terrorism can arise for several reasons. These include when people who are not considered terrorists fall under the scope of a definition, either because the political landscape has changed, political motivations underlie use of the label (Carlile, 2007; Meisels, 2008), or false dichotomies conflate the term with others such as 'freedom fighter' (Ganor, 2010). Difficulties also arise over the breadth of scope (Horgan and Braddock, 2012), the inclusion of glorification and supporting crimes and the importance (and definitions) of sub-terms such as severity, civilian, non-state and political (Schmid, 2011).

Opposition to definitions of terrorism can arise when they capture historical actions that are not conventionally considered terrorism, or where the political landscape has significantly changed. Examples of this include Emily Davison, whose suffragette militancy included stone throwing and arson (Naylor, 2011); former President of South Africa, Nelson Mandela, also co-founder of the militant group uMkhonto we Sizwe (MK); and members of the political wing of the IRA, Sinn Féin, now the second largest party in the Northern Ireland Assembly. When such individuals are defined as terrorists, this leads to an intuitive rejection and opposition to the definition.

The reason for this is, in part, due to the emotional connotations and political impact of labelling terrorism. These influence how the label is liberally applied by politically-

strong parties to their opponents (Meisels, 2008), but also how it is refused by those to whom it is assigned. Conversely, however, some groups welcome being labelled terrorists and it is their opponents who would rather not grant the cachet of respectability or notoriety (Carlile, 2007). The UK's Independent Reviewer of Terrorism Legislation from 2001-2011, Lord Carlile, warns in the context of using UK terrorism legislation to prosecute that the label could upgrade minor, violent or insane criminals, giving them more media time or communication platforms than they should be given. Such concerns also hinder a universally accepted definition. Like other intuitively recognisable concepts, the intuition is based on personal and cultural experiences and is therefore highly subjective. Terrorism is even designed to be subjective (Cronin and Ludes, 2004). Individual beliefs and motivations influencing, even subconsciously, the evaluation of a given terrorism label can cause great difficulties when debating definitions. Not only do labelled groups and governments differ in their opinions on who is a terrorist, but so do people from different world regions, political systems, religions, philosophies, or, more sceptically, with different sources of funding, aid or military protection.

Part of the reason for this is the false dichotomy propagated by careless, or deliberate (Ganor, 2010), discussion, especially in the media, where perpetrators—such as militant suffragettes—are falsely presented as *either* terrorists *or* freedom fighters. These concepts are not mutually exclusive ends of a single continuum, but rather different dimensions. Defining 'terrorists' rather than 'terrorism'—actors rather than the act—can compound this false dichotomy. Many of these difficulties can be avoided by the methodology—more common in academic definitions—of defining the act of violence as terrorism, rather than defining the perpetrator as a terrorist. Saying that a particular group has committed some terrorist 'acts', rather than saying that they are a terrorist group, leaves open both the possibility for them to predominantly engage in non-terrorist activities and for the commentator to have sympathy with their aims—whilst not their tactics. Freedom fighters status is thus independent of whether one's *modus operandi* includes terrorist acts or non-terrorist acts (for example and depending on definition, acts might not qualify as terrorism by targeting only enemy combatants). Moral arguments over the legitimacy of their cause are therefore subsumed into the label 'freedom fighter', rather than the objective assessment of their tactics as 'terrorist

acts'. By aiming to objectively define terrorism as a tactic, issues over the legitimacy of the acts of violence are less problematic in defining terrorism.

One of the first difficulties with defining terrorism is deciding on the correct breadth of definition. Complex definitions, especially academic ones, can be too impractical for operational purposes (Horgan and Braddock, 2012) and the more specific the definition the greater the risk of excluding some acts generally accepted as terrorism. On the other hand, vaguer definitions may receive greater opposition (Horgan and Braddock, 2012) as acts such as violent crime can inadvertently be included under their scope and vague definitions lack the clarity that, for example, a legal definition, requires. Simply deciding on how complex a definition to construct is itself a major challenge.

The scope of a definition is often driven by the number of undesirable activities that a government wishes to fall under the legislation. This can include activities carried out in support of those who commit terrorism, such as financing, preaching or theft of weapons. The UK Government's definition of terrorism includes financing, encouraging, initialising, planning and glorifying terrorism (Terrorism Act 2000). In the US, governmental definitions tend to omit the adjective 'serious', which is frequently used in the UK, EU and some of the UN definitions. While such a technicality may seem minor, the distinction can be important for legal and operational purposes. Consequently, government tend to upgrade crimes to more serious terrorist offences—concerns over this were discussed above (Carlile, 2007)—if they are committed in order to enable a act of terrorism, or for the benefit of a group which undertakes terrorism. Such wide-reaching inclusions rarely occur with academic definitions, which tend to be more focused on the act of violence itself. Upgrading glorifying terrorism to terrorism also raises questions about slippery slope arguments, as does the UK Home Secretary's speech which was in danger of criminalising people for not quite breaking the law (Johnston, 2015). Overall, the want or need of governments to legislate against many activities leads to definitions that cannot be universally accepted.

Definitions also differ in their inclusion of other components. These include the level of premeditation or intention required and, ironically given the origins of the word terrorism, whether the aim has to be to cause terror, or simply have a political objective.

Not only is there debate over which components to include in a definition, there are debates over what those sub-terms—such as 'political reasons', 'civilians' and 'violence'—mean. Such sub-terms are contentious, but fundamental to the definition and each must be explicitly and carefully defined in order to prevent a disagreement over any one of these sub-terms propagating back up to the overarching definition of terrorism. All of the problems with defining terrorism mentioned above—complexity, political implications, subjectivity and moral arguments—repeat themselves for each sub-term.

The most contentious term is '*civilian*' targets (Schmid, 2011). Without such a distinction, guerilla attacks and insurgencies (which international law treats differently) could fall under the definition of terrorism. Including the term, however, might exclude attacks such as the third 9/11 aircraft (American Airlines Flight 77), which was flown into the Pentagon (headquarters of the United States Department of Defense). Also problematic are attacks that are primarily against a military target but where civilian bystanders are caught up—for example the civilian passengers on board Flight 77. The US term is 'non-combatant'. This is interpreted very flexibly depending on what is needed. In a 2005/6 report, 'non-combatants' were defined as

> *"civilians, plus military, whether or not armed or on duty, who are not deployed in a war-zone or war like setting"* (Schmid, 2011).

A 2013 amendment added to the definition:

> *"acts on military installations or armed military persons when a state of military hostility does not exist at that site".*

Schmid also raises the question of whether all those carrying a weapon—hunters, police—are combatants? Whether all members of the uniformed military—chaplains, doctors—are combatants? To an even more extreme definition, Schmid asks to what extent the voting public escape responsibility; the voting public that elects the politicians that order the military attacks. Schmid asks whether there really is an innocent civilian population? Easy counter examples to this, however, are the young children targeted by terrorism no differently to adults. Schmid's survey received responses outlining a range of views on which targets should be included in the definition. Not only is the inclusion or exclusion of non-combatants debated, however, so is the definition of non-combatant. For example, one respondent thought that combatants should be defined as any legitimate fighting force as recognised by the 4th Geneva convention. As Schmid points

out, however, this is dependent on the declaration of war and very few wars are *officially* declared any more (Schmid, 2011). One of the reasons why defining non-combatants is problematic, especially with regards to the status of politicians and off-duty military personnel, is that excluding them in an attempt to define a specific phenomenon can be seen to legitimise attacks against them, since they become excluded from the innocent civilian population.

Reference to '*violence*' also has the potential to include, or omit, for example electronic attacks unless violence is carefully defined. The UK definition chooses to include electronic attacks explicitly. The US definitions do not mention it, but, based on the breadth of their definitions, US officials would probably be quick to label any cyber attack as terrorism. There are also those who argue that electronic attacks should not be considered terrorism at all. That makes three different viewpoints on electronic attacks, one component of violence, which itself is just one term in the definition. Given the number of views on every issue that could arise with every sub-term, it is unsurprising that no universally accepted definition exists.

Another problem occurs with defining of '*political reasons*'. Academic definitions attempting to be as accurate as possible when transcribing terror attacks tend to group all justifications under 'for political goals'. The many motivations that may lead a person to attempt to reach such political goals (religious, philosophical, racial, ethnic etc.) are all included and are irrelevant to the definition so long as there *is a* political intention. Many critics of this, especially with regards to its operational validity for police or security agencies, prefer to distinguish the different motivations people have for carrying out terror attacks, treating religious and political as different possible reasons.

Inclusion of the term '*non-state*' is also problematic, as there is conflict over the moral desire to condemn all evil acts as terrorism, including those by clandestine government agents, versus the necessity of having a very precisely defined term (Schmid, 2011). Thus many academics refer to state-sponsored or state terrorism.

A final difficulty is that the majority of exposure to terrorism is restricted to security agencies and professional intelligence officers, whose work remains classified. Many

experts on the field are therefore locked away in secret and communication between all parties with knowledge to contribute is hindered.

Definitions

There are four main categories of definitions of terrorism: academic, state, media and opponents (Horgan and Braddock, 2012). Although the problems with media and political opponent definitions have been discussed, they are rarely explicitly defined. This section briefly summarises four key state definitions—UK, EU, UN and US—and then introduced a more complex, academic definition that is adopted by this thesis.

UK Definition

The UK Terrorism Act 2000, with amendments in 2006, defines terrorism as follows:

> *"(1) In this Act "terrorism" means the use or threat of action where:*
>
> *(a) the action falls within subsection (2),*
>
> *(b) the use or threat is designed to influence the government [or international governmental organisations] or to intimidate the public or a section of the public and*
>
> *(c) the use or threat is made for the purpose of advancing a political, religious or ideological cause.*
>
> *(2) Action falls within this subsection if it:*
>
> *(a) involves serious violence against a person,*
>
> *(b) involves serious damage to property,*
>
> *(c) endangers a person's life, other than that of the person committing the action,*
>
> *(d) creates a serious risk to the health or safety of the public or a section of the public, or*
>
> *(e) is designed seriously to interfere with or seriously to disrupt an electronic system."*
>
> (The Terrorism Act 2000, 2000; The Terrorism Act 2006, 2006).

EU Definition

The EU defines terrorism as serious criminal offences against people or property that may seriously damage a country or international organisation when committed with the

aim of seriously intimidating... or seriously destabilising or destroying (Council of the European Union, 2002).

UN Definition

The UN and the UN Security Council both have their own definitions. For the Security Council terrorist acts are "criminal acts", including those "against civilians", with the intent to cause death or serious injury (including hostage taking), with the intent to invoke terror and intimidate the public or governments (UN Security Council, 2004).

US Definition(s)

In the US, several governmental organisations have their own definitions. Table 2.1. summarises six—the United States Code, US Code of Federal Regulations, US National Security Strategy, Department of Defense, USA PATRIOT Act and the National Counter-Terrorism (CT) Centre—by identifying the components and sub-terms they include (Definitions of terrorism: United States, Wikipedia).

Table 2.1. Component sub-terms of official United States definitions of terrorism

Definition	Premeditated	Politically motivated*†	Non-combatant target(s)	Sub-national actor(s)	Unlawful acts	Intimidation an objective	Incl. government targets
United States Code	■	■	■	■	■	■	
US Code Federal Regs.		■*			■	■	■
US Nat. Sec. Strategy	■	■	■	■			
Department of Defense		■†			■	■	■
USA PATRIOT Act		■	■		■	■	■
National CT Centre	■	■†	■	■			

The components and sub-terms included in six United States Government definitions of terrorism (Definitions of terrorism: United States, Wikipedia). *socially motivated; †religiously motivated.

An Academic Definition

The following definition constitutes the substantial overlapping features in the definitions adopted by multiple academics (Hoffman, 2004; Meisels, 2008; Laqueur, 2012; Schmid, 2011).

"Non-state[(a)] violence[(b)] against civilians[(c)] for political purposes[(d)]."

(a) Non-state shall exclude those who are employed by or substantially supported by a state. State violence against civilians for political purposes would be defined as war crimes.

(b) Violence will be defined as acts which can cause severe harm meaning that there is the potential of risk to human life beyond the actor's. This includes, but is not limited to, severe physical attacks, severe mental harm, severe economic harm or severe cyber attacks.

(c) *Civilians shall exclude amongst others: government leaders, off-duty military personnel located in war theatres and on-duty military or governmental personnel located outside war theatres. Political, state violence against non-civilians is war, political non-state violence against non-civilians is guerilla warfare or insurgency.*

(d) Political purposes shall mean there is any form of 'relatively' wide scale social, cultural, economic, religious or other change that there is an objective to bring about. The motivations for wanting said change are irrelevant. Examples of non-state violence against civilians without political purpose include, but are not limited to, violent crime only for personal/group financial gain, serial killers seeking seeking widespread public terror only for the increase of their personal notoriety.

The above definition is still far from perfect. Civilian is the most contentious sub-term to define (Schmid, 2011) and a range of minor adjustments to the definition are equally valid. This version excludes attacks against military institutions or political leaders—guerilla attacks—as well as attacks by clandestine government agents. Although there are arguments to include them, they only add to the already problematic heterogeneity—without necessarily being more relevant for analysis of terrorist groups on Twitter. This definition in no way seeks to undermine the severity or reprehensibleness of the specifically excluded acts, however the aim here is a clear, unambiguous definition rather than the moralistic desire to condemn 'all' intuitively terrorist acts within that

term. In order, therefore, to make it clear under what circumstances this work can be integrated with other research, the above definition is the viewpoint from which this thesis considers terrorism.

Practical Definition

In spite of the carefully argued definitions above, in reality, this thesis will adopt a more practical approach. Terrorist Twitter users rarely identify themselves as a member of a proscribed terrorist organisation—only 13% (Wright *et al.*, 2016). Twitter users can also inflate for egotistical reasons, or mask for security reasons, their importance, connections and level of violence. As these cannot be objectively verified, an alternative definition must be adopted.

This scope of this thesis, therefore, is simplified to those who are highly interlinked with other terrorist or extremist accounts. Obviously, this association does not make a person a terrorist. Those actively following and being followed by other terrorists (excluding academics and journalists) are, however, still contributing to the phenomenon of online terrorism and are worth studying—just as the UK Home Office is attempting to crack down on online, extremist material, even if it does not actually break the law (Johnston, 2015). Being interlinked with other terrorists is clearly a circular criterion and thus the definition begins with accounts identified as terrorists by reputable newspapers or other academic papers. The principle of homophily—the tendency of people to associate with others similar to them (McPherson *et al.*, 2001)—has been shown to lead to highly intra-linked communities on Twitter that bias their interactions to other members of the community and share a social identity (Bryden *et al.*, 2011; 2013; Tamburrini *et al.*, 2015). Consequently, it is assumed that terrorists bias the accounts that they follow towards the accounts of other terrorists and that those reciprocally following terrorists would themselves be terrorists or extremists. This assumption is discussed and critiqued further in the research and discussion chapters.

2.2.2. Profiling terrorists

Profiling is contentious in any criminal justice situation. It is misrepresented in the media, has a scientism in popular consciousness which is probably illusory and is disputed for its racialised assumptions. Despite this, profiles and traits associated with

terrorism could be useful; the evidence suggests, however, that they do not exist (Rae, 2012; Silke, 2009). The lack of a profile prevents defining terrorism in the symptom-based style of psychiatric definitions. Further, with regards to the social media work undertaken in this thesis, it means that there are no suggested behavioural or personality features for which to search, or against which to validate a machine model. For governments, the lack of a profile of who becomes a terrorist means that it is more difficult to:

> *"maximise the efficiency of resource allocation, increasing the likelihood of the interception of a terrorist attack"* (Rae, 2012).

'Psychosocial profiling' is the consideration of both the psychological traits and characteristics and the social and environmental factors that may lead to or be more common in people of a certain behaviour—i.e. terrorists in this context. The usual features that tend to feature in profiles and might intuitively appear useful in this case, are gender, race, age, educational level and socio-economic status, amongst others. Elements of personality such as psychopathy and other mental illnesses or insanity are also commonly quoted stereotypes of terrorists. They are summarised by Rae (2012) as *racial-physical, psycho-pathological* and *socio-economic.* The next section of this thesis demonstrates why these profiles are misleading and lacking in evidence.

Five key profiling-dimensions—race, gender, age, education and socio-economic status—have been shown to lack predictive or correlative value (Webber, 2010). The majority of, but by no means all, terrorists have finished college and 20% have a doctorate (Sageman, 2004; Victoroff, 2005). Two early, co-operating, Egyptian terrorist organisations had very different demographic profiles; the educated students and businessmen in the organisation carried out the planning, whilst the uneducated 'foot-soldiers' in the other carried out the attacks (Sageman, 2004). A similar picture emerges with socio-economic status, age, race and, increasingly, gender (Webber, 2010). Whilst many terrorists involved in the current generation of Islamist terrorism may well be of Arab or Muslim identity:

> *"a considerable number are not... [including the] second most lethal terrorist in American history, Timothy McVeigh"* (Rae, 2012).

The growing evidence of Westerners travelling from Western Europe and the United States to fight for Daesh (Figure 2.1.) (McCarthy, 2015) emphasises how the fast-

changing nature of terrorism can mean that research can quickly become out of date. Any profiling attempt must therefore recognise that assessment is context-specific (Rae, 2012; Victoroff, 2005).

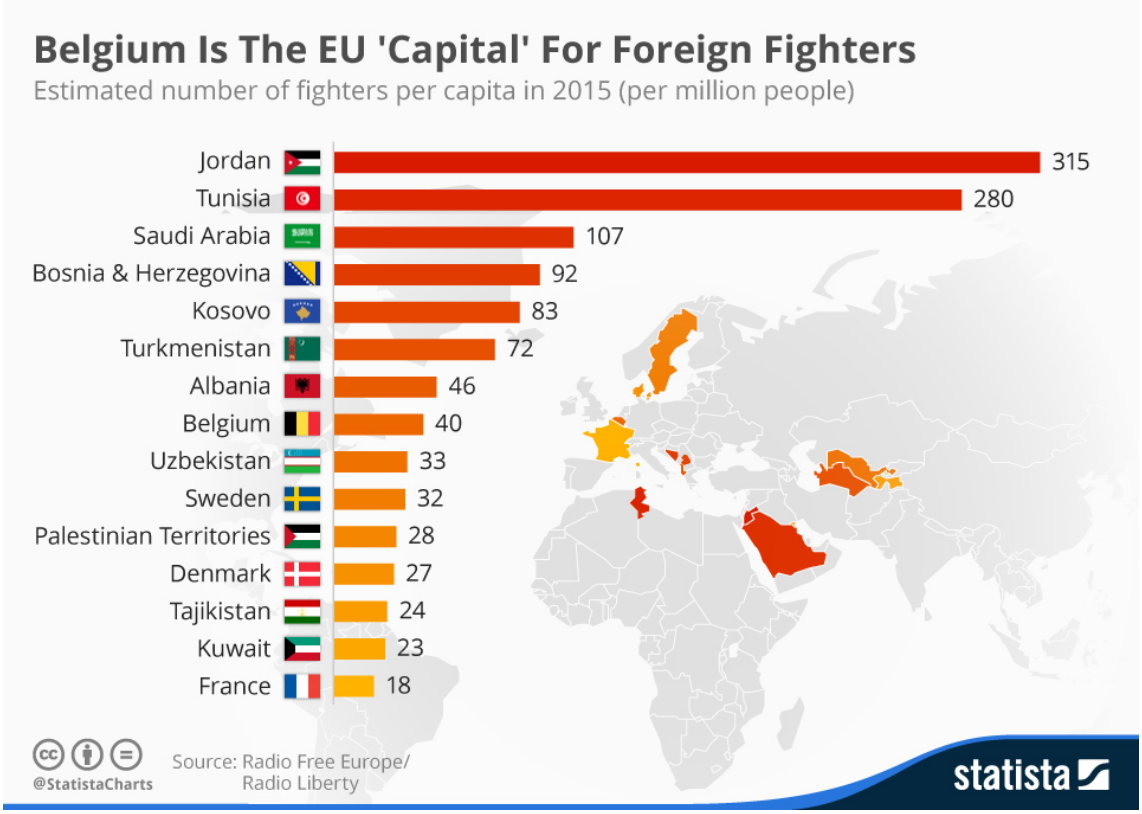


Figure 2.1. Belguim Is The EU 'Capital' For Foreign Fighters. Countries ranked by the estimated number of foreign fighters per capita contributed to the so-called Islamic State in Iraq and Syria in 2015. Reproduced from http://www.statista.com (McCarthy, 2015) under the Creative Commons License CC BY-ND 3.0.

Defining terrorism is, in part, difficult because of heterogeneity in several dimensions (Victoroff, 2005), including *modus operandi* (MO) (e.g. bombings, hijackings, kidnappings, etc.) and motivation (e.g. philosophical, political, racial, religious, ethnic, etc). Considering that serial killers also have a heterogeneous range of MOs and motivations, yet are still fairly successfully profiled, terrorism might also seem ideal for psychosocial profiling. This heterogeneity, however, especially with regards to motivation, should warn against expecting to find a deterministic cause (Webber, 2010).

There is also heterogeneity in the roles carried out by terrorists (Victoroff, 2005). Although terrorists are often stereotyped as both the 'Bin Laden-esque' evil mastermind and the suicide martyr on the ground, a range of other roles are involved. As in any

organization, financiers, accountants, recruiters, technical experts, teachers and logisticians are all needed (Victoroff, 2005). This division of labour was observed in the two co-operating, Egyptian, terrorist organisations described above (Sageman, 2004). Those filling different roles are likely to have more in common with non-terrorist accountants or engineers in their own profession, than with other terrorists in their organisation who fill different roles.

Insanity, Psychopathy, Sociopathy and Mental Illness

Psychopathy is the most commonly assumed trait of terrorists, however reviews have found no differences between terrorists and the rest of the population (Webber, 2010), nor have psychiatrists who have interviewed or sent terrorists questionnaires (McCauley, 2002). So why *do* we tend to assume terrorists must be insane or mad? Other than the fact that psychological profiling is "*widely accepted in both the study of criminology and as a method within law enforcement"* (Rae, 2012), the first reason insanity is assumed of terrorists is the emotional impact attacks have on the affected population. Upon seeing footage, images or reports of a terror attack and the resulting fatalities and casualties, an instinctive response is 'what sort of *sane* person could do this?'. These opinions are frequently voiced in the media and journal publications by non-experts and, causing even more significantly misleading damage, also by those supposed to be experts (Silke, 2009). In a speech to the UK Parliament in 2014, then Prime Minister David Cameron referred to Daesh as "*psychopathic terrorists who want to kill us"* (Morris, 2014).

Silke (2009) also highlights the case of Andreas Baader of the Baader-Meinhoff Gang. Whilst in prison, Baader was the subject of a psychology paper (Cooper, 1978) describing him as:

> *"extremely manipulative, a poseur and something of a pathological liar. Unprincipled and unscrupulous, he entered manhood as little more than an articulate, superficially attractive social parasite... little different from the thousands of psychopaths inhabiting jails and prisons the world over. His time and place – as well as his female associates – gave him the opportunity to display his psychopathy in this distinctive, terroristic fashion"* (Cooper, 1978).

Silke then mentions another paper (Rasch, 1979) by psychiatry professor Wilfried Rasch, published 18 months after Baader's death. In this Baader is described as having no psychotic, neurotic, fanatic, paranoid or psychopathic features. Silke's main point is that whilst Rasch had “extensive, personal contact” and “met and assessed” them while in prison, Cooper had formed his conclusion based on “reading second-hand reports... such as stories from newspapers and magazines” (Silke, 2009).

There have also been significant changes to the nature of terrorism. Prior to the 1980s terrorists were best described as 'lone assassins' (O'Connor, 2004), whereas with the subsequent formation of transnational terrorist organisations: *“as terrorism changed, so did the types of people who became terrorists”* (O'Connor, 2004). O'Connor argues that that may slightly excuse the early obsession with the view that terrorists were mentally ill; perhaps 'lone terrorists' could suffer from mental illness, whereas large organisations are the “*product of other phenomena*” and specifically avoid recruiting the mentally ill for reasons including loyalty, co-operativity, ability to function independently and discreteness (O'Connor, 2004).

Just as the evidence of Westerners travelling to fight for Daesh (McCarthy, 2015) introduced above emphasises how research can quickly become out of date, the increase in (or reversion to) lone wolf attacks over the last few years has led to an emergence of research identifying behaviours and traits (Corner, 2015; Hamm and Spaaj, 2015; Moskalenko and McCauley, 2011):

> *“The odds of a lone-actor terrorist having a mental illness is 13.49 times higher than the odds of a group actor having a mental illness. Lone actors who were mentally ill were 18.07 times more likely to have a spouse or partner who was involved in a wider movement than those without a history of mental illness. Those with a mental illness were more likely to have a proximate upcoming life change, more likely to have been a recent victim of prejudice and experienced proximate and chronic stress.”* (Corner, 2015).

Lone wolf terrorists also sometimes broadcast their intent to attack, often more than once and including on Twitter: 84% of pre-9/11 cases, 76% of post-9/11 lone wolf terrorists (Hamm, 2015). Anecdotally, Anders Breivik, the perpetrator of the Norway

attacks on July 22, 2011, was diagnosed with a psychotic disorder and a personality disorder by two forensic evaluations (Melle, 2013).

Furthermore, significant changes to psychological theory can also explain why terrorists were considered insane. Prior to the 1960s, psychology and psychiatry considered suicide to be an altruistic or "rational response to certain situations"—for example 'going over the top' and Kamikaze pilots were suicide missions for political goals that are not considered markers of insanity (*Silke, 2006 in* Webber, 2010). After the 1960s, however, suicide was considered a pathology that "deviated from a normal response" (*Silke, 2006 in* Webber, 2010). Theories from between 1960 and 1980, after suicide became considered deviant but before the formation of transnational terrorist groups that avoided the mentally ill, supported a model of the terrorist as insane. This argument, however, incorrectly assumes that suicide attacks have been the norm for terrorist attacks since 1960. In summary, although there is no evidence of suicide bombers as insane, this label is a relative one, based on when and who constructed it (O'Connor, 2004; Webber, 2010).

Other traits that might align with psychopathy are also considered. Sensation (Victoroff, 2005) and celebrity-seeking are often cited. Both, however, are common outside terrorism, the former in extreme-sports junkies and members of the armed forces and police. Although most terrorist acts do attempt to affect an audience, there are many cases where no group or individual claims responsibility—some acts are for the audience of God alone (*Silke, 2003 in* Webber, 2010). Sensation and celebrity-seeking are, therefore, neither accurate nor precise indicators of terrorism.

On the other hand, some research suggests that profiles may exist. Psychological profiles of suicide bombers may exist, as they may share traits such as the 'authoritarian personality' and risk factors that increase the probability of suicide (Lester *et al.*, 2004). Work by Ferracuti and Bruno concluded that terrorists have authoritarian-extremist personality, disconnection with reality and ideological "vacuity"; Sullwold concluded that they fall into two categories—'unstable, egoistic and apathetic' and 'intolerant, paranoid and hostile neurotic' (Rae, 2012). Despite, however, these examples often being cited in support of terrorist abnormality, Ferracuti himself only considered terrorists to be "slightly altered, at most" (Silke, 2003).

Profiling summary

As useful as it would be to have a terrorist profile to search for or validate against on Twitter, the evidence suggests that no such profile exists. Terrorism can be committed by anybody. Recognising this could overcome the potential damage of blindly discriminating against people of a certain race or demographic, whilst letting the guilty slip through the net. The US 'Screening of Passengers through Observation Techniques' (SPOT) program let 23 terrorists travel through SPOT points without interception—a 100% failure rate (Rae, 2012). On the other hand, as terrorism continues to evolve and larger volumes of data are collected, a variety of accurate, *context-dependent*, profiles may emerge (Victoroff, 2005). As it stands, there are no characteristic behaviours, profiles or personality traits associated with terrorism for this thesis to rely upon.

2.2.3. The terrorism landscape in 2013

At the beginning of this doctoral research in September 2013, the landscape of terrorism was dominated by a range of groups (Home Office, n.d.). The categories of organisations in Table 2.2, however, make useful distinctions. The first, al-Qaeda (AQ) and subsidiary groups, were heavily weakened and almost erased by the Bush administration's "war on terror". The second were unofficial sister organisations of AQ. These included al-Shabaab, who that very month carried out the 80-hour Westgate Mall attack in Nairobi, Kenya (Howden, 2013), Boko Haram and ISIS. ISIS had only recently renamed itself from ISI and extended the fight from Iraq into Syria (Stern and Berger, 2015). By September 2014, only twelve months later, ISIS had become the dominant player in Islamic terrorism, formally split from al-Qaeda (February 2014) (Home Office, n.d.; Stern, 2015) and declared a Caliphate (June 2014) and changed its name once more, this time to "the Islamic State" (Daesh). Finally, there were unrelated groups with different objectives around the world, although these have formed a small part of international and domestic terrorism in the last decade.

Table 2.2. Categories of terrorist organisations active in 2013.

Category	Abbreviation	Organisation
al-Qaeda	AQC	al-Qaeda central
	AQI	al-Qaeda in Iraq
	AQIM	al-Qaeda in the Islamic Maghreb / Algeria
	AQAP	al-Qaeda in the Arabic Peninsula / Yemen
al-Qaeda linked	HSM	al-Shabaab (Somalia)
	BH	Boko Haram (Nigeria)
	ISIS	Islamic State in Iraq and Syria / Levant (formerly ISI / Islamic State of Iraq October 2006 – April 2013)
Syrian rebels	JAN	Jabhat al-Nusra
	JAA	Jund al-Aqsa
Others	IRA	Irish Republican Army (former Real / New)
	IRA	Irish Republican Army (former Provisional)
	Hamas	Izz al-Din al-Qassem Brigades (Palestine)
	Hizballah	Hizballah Military Wing (Palestine)
	ETA	Euskadi ta Askatasuna (Basque Spain)

This table lists terrorist organisations, active in late 2013, grouped by common association. These groups are proscribed by the UK Home Office (Home Office, n.d.). Although a selection of the largest and most infamous organisations from the Home Office's wider list are given here, this table is far from complete.

Terrorism on Twitter in 2013

Large volumes of representative data from online social networks have made previously hard to study social phenomena amenable to investigation (Danescu-Niculescu-Mizil *et al.*, 2011). This also appears to be the case within the apparently unique domain of terrorism (Lynch *et al.*, 2014). It was not, however, a foregone conclusion at the outset of the work in this thesis—2013—that this should be the case and that terrorists should be active on Twitter. There were several arguments why they should not have been, principal amongst which was that Twitter is a publicly visible site. Given the lengths that members of terrorist organisations go to to conceal their identities and activities, at first glance it might seem surprising and foolish for them to use a public site.

Klausen *et al.* (2012), however, gave several reasons to support the hypothesised existence of Twitter accounts belonging to, or supporting, terrorists. First amongst these was the ease with which illegal or extremist material could be concealed amongst the mass of online postings (Klausen *et al.*, 2012). Secondly, even when posts are not concealed, there is protection under the First Amendment for material on sites legally based in the US, including Twitter, regarding incitement and hate speech (Klausen *et al.*, 2012). At the time of Klausen's writing and indeed for several years after, Twitter's policy and procedure for removing material was more "lengthy and restrictive" (Stern and Berger, 2015) than public pressure later forced it to become. Thirdly, the possible alternative locations for terrorists to communicate online, seen in the historical models of online communication

> *"after September 11, the message boards... more commonly referred to as online forums or just "the forums"... became the preferred social networking tool for jihadists"* (Stern and Berger, 2015, pp.128)

suffer two major flaws. Firstly, they limit access to potential recruits, whereas anyone can create a Twitter account and find and follow interesting communities. Secondly, dedicated forums are vulnerable to complete annihilation as internet hosting servers are often based in the US and happy to comply with government closure requests. With Twitter, however, the legal shield given to mainstream platforms compels law enforcement and service providers to close down accounts or remove videos on a one by one, *ad hoc* basis (Klausen *et al.*, 2012). Finally, there is historical evidence that *"terrorists follow the same technological trends that everyone else does"* (Stern and Berger, 2015, pp.130), adopting videotape, email newsletters and digital video in the 1980s, dial-up bulletin boards, chat rooms and online forums in the 1990s, all at *"around the same time early-adopting consumers did"* (Stern and Berger, 2015, pp.128).

As a side note, by the latter stages of this thesis work (winter 2015—summer 2016), the cycle has already begun to repeat itself; an increasing number of jihadists and their supporters are migrating to the encrypted messaging service Telegram (https://telegram.org) (Bunzel, 2016; Weimann, 2016). Alongside this we see the usual resistance from the last vestiges of support for the status quo:

> *"a jihadi author is lamenting the decline of the social media platforms, warning users against migrating to Telegram, an encrypted messaging service and calling for the revival of Twitter and Facebook"* (Bunzel, 2016).

The counter-argument that terrorists would not use a publicly available site in the way that they adopted those more private technologies incorrectly conflates the reasons they use social media; not all of their activities involve plotting attacks. The question, therefore, becomes, 'for what purposes or functions do they use Twitter'? Terrorists use Twitter for a variety of reasons similar to their use of websites and the internet in general (Qin, 2011): spreading their messages to a wide audience, recruitment, indoctrinating further those drawn to them—like a crucible of radicalisation and finally, (although not comprehensively) for seemingly mundane conversation amongst friends (Stern and Berger, 2015; Wright *et al.*, 2016). Those behind the account do not necessarily see all of these activities as terrorism, or illegal. Political campaigning, religious preaching, recruiting are all safe activities—relative to plotting attacks—to be carried out in public, although it could be argued that it may still draw the attention of intelligence and security services, onto whose watchlists they could be placed. Furthermore, additionally countering the 'they would not be on a public site' argument, are the range of easily achieved security options. Firstly, they could just protect their accounts; make their content private rather than publicly viewable, or just talk in private messages. Secondly, they could completely mask their intention by speaking in codes and using unrelated names, pictures and content. Or, finally, even whilst posting blatantly extremist content, users can conceal their identity with a fake email address and using a proxy and/or TOR (The Onion Router). Irrespective of which, if any, tactics they choose, we are likely to only see material that they consider unclassified.

Despite these logical arguments, perhaps the most simple reason to look for terrorists on Twitter is evidence that they are there. At the outset of this work a range of newspaper and think tank reports revealed the presence of a variety of al-Shabaab, al-Qaeda and ISIS accounts. The plotters for the foiled 2013 attack on the Myanmar embassy in Jakarta met, chatted and planned on Facebook (abc.net, 2013). Furthermore,

> *"the suppression of [al-Qaeda's top-tier forum Shamukh al-Islam]… accelerated an already growing trend: the migration of jihadi propaganda from*

web forums to social media. In response to the blackout, many jihadi groups, media outlets and individuals created new accounts on Twitter" (Zelin, 2013).

As the offline terrorism landscape evolved over the course of this thesis work, so did the online landscape. This is discussed in the 2016 section of the literature review.

In conclusion, terrorists and their supporters are present on Twitter, albeit for specific reasons. A more significant problem is not discovering whether they are present, but identifying the correct accounts through the noise of interconnected supporters, journalists, academics and curious members of the public. This means that there is a lot of traffic data available to researchers, investigators and experts, but it is noisy and the signal is weak (Brynielsson *et al.*, 2012). The aim of this research is to spot meaningful and predictive patterns amongst that social media noise; to investigate how best to analyse the traffic and tune in to terrorist signals.

Specific accounts on Twitter

The following paragraphs describe the landscape of terrorism on Twitter as of the first few months of research on this thesis (up to February 2014). Multiple terrorist organisations had a presence on Twitter. Three prominent examples are *al-Qaeda*, a*l-Shabaab (HSM)* and the *Islamic State of Iraq and the Levant (ISIS)*. There are also accounts that are potentially connected to these accounts behind these scenes, as well as fake/hoax accounts which impersonate the real organizations and cause confusion, anger and can hamper the authorities.

al-Shabaab

al-Shabaab is a Somalia based organisation, officially designated as a terrorist organisation by eight countries. In 2012 it declared allegiance to al-Qaeda and although there have been some subsequent defections and appeals to switch allegiance to Daesh, it remains a supporter of al-Qaeda. Their official name is Harakat al-Shabaab al-Mujahideen (HSM) which translates as *Mujahideen Youth Movement.*

On Twitter, one of their early accounts, active from December 2011 until January 2013, was HSM Press Office—*@HSMPress* (Hudson, 2012; Barnett, 2013). Their subsequent accounts followed a play on the letters 'HSM' and the word 'press' in their handles and

most took the name "HSM Press Office"*: @HSMPress1* (Barnett, 2013; Straziuso, 2013); *@HSM_Press* (Barnett, 2013); *@HSM_PressOffice* (Barnett, 2013); *@HSMPrOffice* (Barnett, 2013); *@HSM_Pr* (Barnett, 2013); *@HSM_Info* (Mohamed, 2013).

In addition to the naming consistency, all ten official accounts used almost the same profile picture. This banner or flag is known as al-rāya (the banner), rayat al-ʻuqab (banner of the eagle), or simply, the "*black banner*", "*black standard*", or "*black flag of jihad*" (Black Standard, Wikipedia) and has two main uses. Firstly, it is the "war flag" of Al-Shabaab, however it is also the flag of the Islamic State of Iraq and the Levant (al-Qaeda in Iraq). The upper part of the flag is the Tawhid portion of the Shahada in white Arabic script on a black background—it reads: "*There is no god but Allah"* and is an Islamic creed declaring belief in the oneness of God and the acceptance of Muhammad as God's prophet. The lower symbol on the flag is the *Seal of Muhammad*. This reads *"Mohammed Messenger of God"*.

Although all official accounts used this flag, the exact version of the image differed. Furthermore, there was variation in the use of the black and white flag (x2) or a black and grey version (x5). A second version of the flag also exists, with both colours reversed (i.e. black text on a white background. Rather than the war flag, this version is the administration flag, however all of the al-Shabaab accounts used the war version of this image.

There are also a selection of hoax accounts and some of uncertain authenticity (Barnett, 2013). During the Westgate Mall attack (Howden, 2013), as official HSM accounts were suspended, seemingly identical—but fake—accounts popped up (Barnett, 2013). These accounts *"claimed to have the names of the Shabaab members carrying out the Westgate attack"* (Barnett, 2013), but an official al-Shabaab spokesperson refuted this, saying *"the account that published the names is fake [and] the names are fake"* (Barnett, 2013).

A subsequent review revealed subtle differences between the fake and official accounts, although the fakes still follow many of the patterns exhibited by the official accounts: *@HSM_Official1; @HSM_Press2* (Barnett, 2013); *@HSM_PresOffice2* (Barnett,

2013); *@HSMPress_; @HSMPRESS99; @HSMPress4;* @HSMPressoffice2; *@HsmInfo1*.

al-Qaeda Forums/Websites

Other types of al-Qaeda accounts repeatedly cropped up on Twitter as of September 2013. The first type appears to be a more generic al-Qaeda account, possibly related to central leadership and related to a specific, official, al-Qaeda website and forum—Shamukh al-Islam (Gertz, 2013): *@Shomokhalislam* (Gertz, 2013; Gorman, 2013); *@Warshashamikh*. These accounts were either less successful, or less prepared, to continue creating new accounts following suspension and are thus less active as of 2016.

With these accounts a similar naming pattern emerged, although the handles showed slightly greater variation than those of al-Shabaab. In addition to both accounts using the same profile picture, these accounts consistently using the same background images.

Early ISIS / ISIL / (Pre-Islamic State / Daesh)

ISIS created its first official Twitter account in October 2013, launched the "Dawn of Glad Tidings" Twitter app in April 2014 that was capable of sending tens of thousands of tweets per day and then had it terminated after spamming World Cup hashtags with graphic images of executions in June 2014 (Stern and Berger, 2015).

It's first account was *@e3tesimo*—named مؤسسة الاعتصام. This was *"an official "media foundation" under the name al I'tisaamm, an Arabic reference to maintaining Islamic traditions without deviation"* (Stern and Berger, 2015, pp.143).

Although Twitter suspended *@e3tesimo "in late 2013 or early 2014, for reasons that were not entirely clear"* (Stern, 2015, pp.154), subsequent accounts followed the same patterns as other organisations, playing on the letters 'e3tsm' in their handles and most took the name مؤسســة الاعتصــام: *@e3tsemo*; *@e3tasimo*; *@e3tsm*; *@e3tsemo_*; *@e3tesamo*; *@wa3tasimo*; *@wa3tasimu* (Stern and Berger, 2015). Also as before, there was consistency in profile images. 5 used the administration version of the war flag described above, with only *@e3tsemo_* using the war flag.

Others

Other groups recognised in the media as having official media accounts present on Twitter as of 2013 are listed in Table 2.3.

Table 2.3. Terrorist groups with a presence on Twitter as of 2013

Group	Description and *@Twitter_accounts*
al-Qaeda in the Islamic Maghreb (AQIM)	section of al-Qaeda in North Africa; the Abdullah Azzam Brigades (AAB), an al-Qaeda militant group with networks in Egypt, Iraq, Syria, Jordan, Gaza and Lebanon: *@azzambrigades_*; and the AQIM media arm: *@Andalus_Media* (Agence France Presse, 2013);
Syrian Electronic Army (SEA)	a non-state collection of hackers aligned with the Syrian Assad government the group has had numerous successful cyber attacks on the websites and Twitter accounts of western news agencies, software companies and other (usually western) organisations—often with the intention of causing economic damage or stock-market dips such as that of April 23, 2013: *@Official_SEA; @Official_SEA16; @SEA_Official17; @SEAWebsite4; @SEATh3Pr0; @Th3Shad0w_SEA; @Vict0rSEA; @sea_the_soul; @SEA_Leaks6;*
Ezzedeen Al Qassam Brigades (Izz ad-Din al-Qassam)	the military wing of Hamas—a multi-celled terrorist organization based in Gaza to defend Palestine against the Israelis. Their stated objectives are to "evoke the spirit of Jihad", "defend Palestine" and "liberate Palestine": *@AlqassamBrigade* (Hudson, 2012; Agence France Presse, 2014; Dvorin, 2014); *@qassamBrigades* (Dvorin, 2014); *@qassamBrigade* (Ynet, 2014; Dvorin, 2014); *@qasamBrigade* (Dvorin, 2014); *@qassam_arabic* (Agence France Presse, 2014); *@qassamhebrew* (Agence France Presse, 2014); *@spokespers; @qassamfeed*;
al-Nusra	previously allied with ISIS, then a branch of al-Qaeda in Syria and subsequently an independent rebel alliance
Hezbollah	translated as "Party of God", an Islamic group based in Lebanon. They have a strongly affiliated satellite television station known

	as al-Manar and a radio station al-Nour The US state department considers these entities and their parent company the Lebanese Media Group, to be the media arm of Hezbollah (U.S. Department of the Treasury, 2006): *@almanarnews* (Williams, 2011; Hudson, 2012); *@AlmanarEnglish; @almanarFrench*;

Organiser Accounts

There are also some accounts of interest because of their associations with the above accounts. These accounts tend to demonstrate three key features that may prove to be of interest. Firstly, they reciprocally acknowledge the above accounts via reciprocal following. Secondly, this mutual link is either exclusive or occurs very early on into the life of the newest account. Finally they have similar or connected profile pictures. These three features could suggest that they have some form of foreknowledge either by knowing, or being, the person responsible for setting up the new account.

Example one: *@ajnad_*. While *@e3tesimo* was active, the only account it followed was *@ajnad_*, despite having many followers that it could potentially have reciprocated. *@ajnad_* is also the only reciprocated follower out of 20 for the account *@e3tsm*. Furthermore, although *@e3tesimo* was only up for around 5 hours, *@ajnad_* was already a reciprocal follower very early on into the life of the former account. *@e3tsemo_* followed 111 users, but the first account it followed was *@ajnad_*. There is also profile picture similarity. *@ajnad_* uses the war flag described above, whereas the ISIS accounts use the administration version.

Example two: *@s7bhijratain*. *@s7bhijratain* demonstrated foreknowledge of the al-Shabaab account *@HSM_Info*. At 19:27 on 15th December 2013, they retweeted a tweet from *Live From Mogadishu (@Daudoo)*:

> *"BREAKING:According to Sources close to Al-Shabaab,the group is expected to relaunch its official twitter account soon.#Somalia #Shabaab #HSM"*

Ten minutes later at 19:37 they made their own tweet:

> *"#PRT I can confirm that the official handle of #HSM is expected to back in the coming days bi'ithnillah. Sit tight ;)"*

The following day, the new HSM account *@HSM_Info* went live on Twitter, apparently confirming their connection. *@s7bhijratain* also mutually followed several of the HSM accounts and uses the same profile picture as the HSM accounts (in black and grey).

Example three: *@Vict0rSEA* and *@Official_SEA16.* The first account that each follows is the other and they both display variations on the same SEA logo.

Summary

In summary, the similarities in name, reciprocal following and profile pictures, as well as in language, are potentially enlightening. The real accounts remain consistent, with the HSM ones using good, journalistic English at the level of a native speaker and beginning their life with a blessing in Arabic. The hoax HSM accounts tend to use informal English, spelling and punctuation errors and overuse exclamation marks. The al-Qaeda accounts speak either entirely in Arabic or entirely in very bad English, indicative of a non-native speaker. The *@e3tsemo* accounts use almost entirely Arabic. Similarities in profile features between accounts have also been observed in other work (Stern and Berger, 2015). Such observations will play a crucial role in informing many of the hypotheses and machine methods developed throughout this thesis in order to tune in to meaningful signals.

2.2.4. The terrorism landscape in 2016

The previous section described what was known of the terrorism landscape at the outset of this doctoral work. During the course of the research for this thesis alone, between September 2013 and May 2015, there have been at least 61 reported terrorist attacks (List of terrorist incidents linked to ISIL, Wikipedia). The targets of these attacks include the Westgate Mall in Kenya; the Jewish Museum in Brussels, Belgium; stabbings in Melbourne, Australia; shootings at the Ottawa Parliament, Canada; a hatchet attack on the New York subway, America; a café in Sydney, Australia; Dijon, France; Charlie Hebdo magazine, France; Copenhagen, Denmark; and a convention in Texas, America. Far from diminishing with either the long running "war on terror", or the withdrawal of troops from Iraq from 2007 through until 2011, the threat from terrorism appears to continue growing. Understanding and mitigating the threat from terrorism, therefore, is a key aim of this thesis.

Daesh

At the outset of this work, ISIS had only recently renamed itself from ISI and extended the fight from Iraq into Syria (Stern, 2015). By September 2014, only twelve months later, ISIS had become the dominant player in Islamic terrorism, committing 2,494 (11.5%) of the attacks during 2014 and 2015 (National Consortium for the Study of Terrorism and Responses to Terrorism (START), 2016).

It had also formally split from al-Qaeda in February 2014 (Home Office, n.d.; Stern and Berger, 2015), declared a Caliphate in June 2014 and changed its name to "the Islamic State" (IS). There is, however, discussion on how to refer to the group. Some think that "Islamic State" is "the most correct usage" (Stern and Berger, 2015), but accept that journalists have embraced "ISIS". Others, such as the French foreign minister, Laurent Fabius, have suggested that we reject their self-designation

> *"This is a terrorist group and not a state... The term Islamic State... blurs the lines between Islam, Muslims and Islamists"* (Shariatmadari, 2014).

Thus "Daesh", the acronym of the Arabic phrase *al-Dawla al-Islamiya fi Iraq wal-Sham* is recommended.

> *"Although that too means The Islamic State in Iraq and al-Sham, to non-Arabic speakers it is just a noise. Free of associations at the moment, it will become infused with our ideas about Isis—just as al-Qaida, rather than bringing to mind its translation, "the base", has become irrevocably linked with death and destruction"*(Shariatmadari, 2014).

In this thesis, the name Daesh will be adopted in place of ISIS / ISIL / IS / Islamic State wherever possible. When quoting or referencing other publications, however, this might not always be possible.

The Baqiya family

As the offline terrorism landscape evolved over the course of this thesis work, so did the online landscape. From the mixed al-Qaeda and ISIS community of groups described above, the Daesh supporting Baqiya family has emerged (Amarasingam, 2015; Huey and Witmer, 2016; Miller, 2015). The largely bottom-up Baqiya family is:

> *"a loose network of Islamic State supporters from around the world who share news, develop close friendships, and help each other"* (Amarasingam, 2015).

Baqiya comes from their remaining part of their slogan, "baqiya wa tatamaddad" (remaining and expanding) (Zelin, 2015) and is often a marker of their accounts via a "Baqiya Shoutout" (Berger, 2016). This Twitter family of openly-communicating, accepted members of the Daesh online community of terrorists, extremists and their supporters is large, generating an enormous volume of traffic. Berger and Morgan (2015) estimate—conservatively they say—that "ISIS supporters" used over 46,000 accounts in only four months from September to December 2014. In three years, plentiful evidence has emerged of this different, yet more widespread, terrorist presence: academic evidence (Amarasingam, 2015; Berger and Morgan, 2015; 2016; Bodine-Baron *et al.*, 2016; Huey and Witmer, 2016; Klausen, 2015; Magdy *et al.*, 2015; 2016, Wright *et al.*, 2016), journalistic evidence (Agence France Presse, 2014; Arthur, 2014; Barnett, 2013) and governmental evidence (Jones, 2014). In 2014, the new Director of GCHQ said that

> *"they [US technology companies such as Facebook and Twitter] have become the command and control networks of choice for terrorists and criminals, who find their services as transformational as the rest of us"* (Jones, 2014).

Summary

In summary, the advent of social media led to a substantial volume of terrorist signals amenable to analysis. Observations that accounts have similarities in name, reciprocal following, profile pictures and language inform the investigation of this phenomenon throughout this thesis.

2.3. Language

The second area of background literature informing this thesis is language research. Language is important to this thesis as it provides the model system and proxy to uncover novel patterns used in each study. The following section of the literature reviews considers both the literature on language as a proxy for associated sociological and behavioural characteristics and our biological and psychological / linguistical understanding of how communication can lead to language change and evolution.

2.3.1. Definitions

Table 2.4. Definition of key terms for language evolution as used in this thesis

Term	Definition
Evolution	descent with modification (Darwin, 1859; Pagel *et al.*, 2007); change over time via the inheritance of changes
Heritability/ Inheritance	the percentage of variation in a trait that is transmitted to offspring, whatever the mechanism of transmission (Danchin *et al.*, 2011); the percentage of offspring variation in a trait that can be explained by progenitor variation in the trait (Falconer and Mackay, 1995)
Alice and Bob	adopting the convention of computer science, cryptography and physics (Schneier, 1994), these names will be used for prosaic convenience in place of 'person one' and 'person two' during complicated explanations.

2.3.2. Information conveyed

In this thesis, analysis of language is used to tune in to patterns and changes in terrorist signals. This is possible as language conveys information about the person producing it. As well as 'what' is being said—the semantic content—which can be too subjective or ambiguous for computers to interpret, 'how' it is said—the linguistic style—can also capture information (Pennebaker *et al.*, 2003). Geographical information can be revealed by 'the' language (e.g. English, French, German etc.) that a person uses and if the language has gone extinct, analysis of language in historical texts this can reveal temporal information through . On a smaller scale, people's most commonly used

words and word choices can hold information on their age (Pennebaker *et al.*, 2003), social demographic, regional location, personality (Mairesse *et al.*, 2007), or emotional state (Mohammad and Yang, 2013). Sixteen language features have gender effects (Pennebaker, 2003). Punctuation and spelling can also indicate educational level:

> *"populist guides to grammar, spelling, and punctuation... support the idea that linguistic '(in)correctness' or '(in) competence' can and even should be used as a primary index for... education, intelligence, maturity, trustworthiness and habits"* (Hardaker, 2013).

Average word length can predict social group (Bryden *et al.*, 2013) and social power and status (Tchokni *et al.*, 2014). Social power is also predicted by emoticon use, as well deviating less from ones usual word choice or linguistic style (Tchokni *et al.*, 2014). Even more subtle features, such as three-letter word-endings, have been shown to be good predictors of group identity (Bryden *et al.*, 2013). Whilst the weak, often correlative nature of these association means that a degree of caution must be observed, language features are a viable proxy for underlying patterns. There is an assumption in looking for shared language patterns and changes, that any patterns have arisen via cultural transmission and inheritance of information.

2.3.3. Cultural inheritance

Evolution at its simplest is "descent with modification" (Darwin, 1859; Pagel *et al.*, 2007). The Modern Evolutionary Synthesis (MS) is our current understanding of evolution and its mechanisms, integrating theory from all areas of biology. MS assumes that the *thing*—and the only thing—that descends with modification is the genome (Bonduriansky *et al.*, 2009; 2012).

There is, however, a growing body of work calling for an Extended Evolutionary Synthesis (EES) (Bonduriansky *et al.*, 2009; Danchin *et al.*, 2010; 2011; Jablonka *et al.*, 1995; 2005). The principle around which such an updated synthesis revolves is that there are multiple forms of inheritance. As well as the MS genetic inheritance, a growing body of evidence suggests the possibility of epigenetic inheritance (Carone *et al.*, 2010; Jimenez-Chillaron *et al.*, 2009; Ng *et al.*, 2010; Radford *et al.*, 2014), maternal inheritance (Birky, 1995) and cultural inheritance (Kirby *et al.*, 2008). Whereas MS defines heritability as:

> *"the percentage of variation in a trait that is genetically transmitted to offspring"* (Danchin, 2011),

Danchin (2011) advocates a concept of 'inclusive heritability', defined as:

> "*the percentage of variation in a trait that is transmitted to offspring, whatever the mechanism of transmission"*.

Inclusive heritability is therefore the sum of its genetic, epigenetic, maternal, cultural and perhaps other, components.

Although the literature assumes that certain behavioural traits, such as the social behaviours considered in this thesis, are transmitted to offspring through cultural inheritance (and even standard quantitative genetics carefully controls for the variance in a trait caused by sharing a cultural environment with the previous generation) cultural inheritance is particularly challenging to prove and quantify. Animal behaviour models of estimating cultural heritability, for example, are hindered by the fact that social imprinting often occurs very early in life, thus conflating genetic and cultural inheritance (Danchin, 2011). Language, however, is an easily quantifiable trait where social learning continues to occur throughout life (Bloomfield, 1933). It is, therefore, an important model system for cultural evolution.

2.3.4. Language evolution

The term 'language evolution' is applied to two different classes of phenomena. Firstly, language evolution can mean the 'emergence' of language, i.e. its evolution from non-existence via the emergence of the biological, genetic, neural and physiological capacity for language (Hauser *et al.*, 2002; Longa, 2013; Mesoudi *et al.*, 2011). There is overlap, due to the similarities in the capacities needed, with the investigation in infants of innate mechanisms such as grammar rule acquisition, whether this is from a biological, linguistic, Chomskyan, or psychological perspective (Dodd *et al.*, 2003; Storkel, 2001). None of these forms of emergent language evolution are the subject of the work in this thesis.

The second meaning of 'language evolution' is language change (Croft, 2006; Pagel, 2009). The course of human history and evolution is closely interlinked with the changing use of language. Language is thought to evolve through the passing of

language features from one individual to another and we adopt language features in a variety of different ways. This phenomenon, therefore, is more information based and is analogous to the study of the mathematics and patterns behind genetic inheritance, transmission and evolution, albeit not necessarily driven by a genetic mechanism. To accurately model the evolution and spread of language it is important to understand the different ways in which we transmit and adopt language features.

There are three areas of language evolution literature that are important for the work in this thesis: studies of vertical change at the population level; mechanisms of vertical change at the level of the individual person; and population level evidence of horizontal changes. (The "Convergence / alignment / imitation / mimicry" section (Section 2.3.6) will also discuss the psychological and sociolinguistic evidence of horizontal transmission).

Firstly, there are historical and comparative linguistics that show us that language changes across periods of history and even from year to year. These macro level studies (Dunn *et al.*, 2005; Gray *et al.*, 2009; Pagel *et al.*, 2007; 2009) use tools from the genetics analysis toolbox to, for example, construct phylogenies of languages to calculate relatedness, common ancestors, point of divergence, or rate of evolution. For example, there is consensus on the relationships between the languages, with well-documented language families such as Romance (Spanish, French and Romanian) and Germanic (German, Dutch and English) (Pagel, 2009). Evolutionary biology has also informed this work, for example, geographic separation can lead to speciation of languages. Prior to modern global transport, language relatedness was highly correlated with the geographic proximity of the speakers. Pagel (2009) also gives the example of the English loss to the French at the battle of Hastings in 1066, after which the English (Germanic) language was bombarded with French words of Romance origin. These approaches assume that the mechanism by which language change occurs is through interaction with the language of others, that language is transmissible at the individual level and that this transmission then has a lasting effect on the recipient's language.

Secondly, there is a body of literature investigating those individual level mechanisms (Castellano *et al.*, 2009; Nowak *et al.*, 2000). This primarily focuses on the easily observed vertical transmission, whereby words are acquired from parents and

immediate family members in the first few years of life—forming an internal information store, analogous to the genome, that can descend from one generation to the next. Although theoretical work has demonstrated that it is plausible that an accumulation of changes at the individual level can lead to macro evolution (Castellano *et al.*, 2009), at the individual level there is only empirical evidence that vertically transmitted language is inherited in a lasting way.

The third body of literature is smaller, but presents population level evidence that neologisms can emerge and spread—following various transmission laws (Eisenstein *et al.*, 2012; 2014; Nerbonne, 2010; Trudgill, 1974)—in time frames incommensurable with vertical transmission (Eisenstein *et al.*, 2012; 2014; Grieve *et al.*, 2016).

In summary, there is plentiful evidence that language evolves. Neologisms spread through communities, changing word usage. The words used across different historical periods have also changed. One of the aims in this thesis aims is to provide empirical evidence of a mechanism of horizontal inheritance of language. The population level evidence of horizontal changes suggest such a mechanism exists, but this assumption has not been empirically validated.

2.3.5. Internalisation / information store

A process by which lasting changes to a person's language can be inherited requires some form of internal storage of language information—a template that influences the words we choose and how often we use them. There are clear parallels here with the internal storage of genetic inheritance—the genome. Just like the genome is a template that influences observable phenotype, each word in a language is a trait and the frequency with which a person uses that word is their value of that trait. The internalised language template can be modelled like a multi-set: a bag in which members can appear more than once. This internalised distribution is then formed depending on the frequency of words that an individual is exposed to and then their language production is based, in part, on that internalised distribution. A person's phenotype is therefore easily observed through their writing or speech. Conceptualising the language information template—which this thesis terms the *lexome*—in this way enables tools from the genetic analysis toolbox to be used.

The important function of this internal language template is that it is updated depending on the frequency of words to which the individual is exposed. 'Internalisation' is a psychological and sociological process where a behaviour that an individual is exposed to becomes incorporated into their nature or normal repertoire of behaviours. Importantly, the behaviour in question can then be observed at other times and in other contexts different to the one in which it first occurred. To show that language is heritable, it is important to demonstrate lasting changes to the internal store of language information, rather that just imitation—which will be discussed in the next section.

2.3.6. Convergence / alignment / imitation / mimicry

A body of literature from the psychology and sociolinguistics communities has looked at how language changes at the level of the individual. This change is often studied as a phenomenon in its own right, rather than being considered as a mechanism by which population or evolutionary changes are occurring. Many previous studies have shown that when people interact, they converge to each other's behaviours (Bonin *et al.*, 2013; Brennan, 1996; De Looze *et al.*, 2011, 2014; Hemphill and Otterbacher, 2012; Iwata and Watanabe, 2013; Mitchell, 2012; Nenkova *et al.*, 2008; Pardo *et al.*, 2006, 2012; Stenchikova *et al.*, 2007; Ward, 2007, Włodarczak, 2013). Other words used in the literature to describe investigations into this convergence include mimicry, compromise, copying, alignment, accommodation and entrainment; all meaning that the behaviour of two or more users becomes more similar over time. Convergence includes conforming to norms such as the fashion of a group with which a person identifies, but also occurs during dyadic (between two people) interactions. Examples of the latter include mimicking the customary greeting of your partner (handshakes, bowing, cheek-kissing etc.), or compromising on the topics of interest or words used (talking to children versus adults, foreigners versus friends, perceived expertise etc. (Jucks *et al.*, 2008)). Convergent accommodation is an immediate and necessary reaction to increase trust and empathy between participants. It has been demonstrated that interactions with more convergence are associated with greater trust and empathy between participants and greater success in whatever task the interaction was occurring to enable (Christopherson, 2011; Hemphill and Otterbacher, 2012; Mitchell, 2012; Nenkova *et al.*, 2008; Steinhauser *et al.*, 2011; Ward, 2007).

Behavioural convergence has also been demonstrated with peoples' language behaviours during conversations. *Within* pairwise interactions, *during* that interaction, people's language becomes more similar to their partner's (Bonin *et al.*, 2013; Brennan, 1996; Danescu-Niculescu-Mizil *et al.*, 2011; Ward, 2007). People imitate one another syntactically (Hemphill and Otterbacher, 2012) and semantically (Branigan *et al.*, 2011; Brennan, 1996; Mitchell, 2012). This imitation is known in linguistics as "Communication Accommodation Theory" (CAT).

In summary, there is widespread evidence that people imitate one another's language. This is, therefore, evidence of horizontal language transmission. Many of these, however, are based on small scale studies (Danescu-Niculescu-Mizil *et al.*, 2011) and all of this work has considered what happens *during* the interaction between people to the convergence that occurs *within* the interaction. Even the study of Internet Movie Database (IMDb) reviews (Hemphill and Otterbacher, 2012) only demonstrates alignment to and imitation of, review norms. It has not been shown whether all of these examples of phenotypic convergence are just superficial, temporary, success-promoting imitation, or whether these changes are inherited and internalised to the lexome in a lasting way—i.e. true descent with modification.

2.3.7. Heritability and confounding factors

Returning to the mathematical principles underlying the definition of heritability (Falconer and Mackay, 1995), the heritability of words is defined by this thesis as '*the proportion of variation in word usage that is the result of transmission'.* In order to demonstrate that language is heritable, it is necessary to show that that proportion is significantly non-zero. Unlike genetic heritability, where each generation of humans corresponds to one generation of the genome, the language template goes through multiple generations within a given human's lifespan. To investigate the variation between subsequent generations of the language template, progenitor template and offspring template must be defined. Progenitor and offspring language templates are the templates at any two time points separated by a cultural interaction (e.g. a conversation), at which descent with modification could occur. When two individual have such a cultural interaction, two new offspring are created. If the progenitor

language templates belong to Alice and Bob, then one of the offspring templates has Alice as the predominant progenitor (conceptually, the maternal progenitor, but with over 99% relatedness) and Bob as the minor progenitor (paternal). The other has Bob as the main (maternal) progenitor and Alice as the minor. Irrespective of how small the contribution made by the minor progenitor, there is potential for cultural transmission to occur.

Returning to '*the proportion of variation in word usage that is the result of transmission',* this can be calculated by comparing word usage in the offspring language against that in the progenitor language. The variation observed, however, will be the sum of several potentially confounding sources of variation (Falconer, 1995) which must eliminated and controlled for. Fortunately, the techniques of standard quantitative genetics (SQG) (Falconer and Mackay, 1995) are easily transferable to this context. Four of the most significant confounding sources of variation are mutation/experimental bias, assortative mating, maternal effects and mimicry. To control for mutation and/or bias, results are compared with the regression of phenotypes of two unrelated people from different environments. The work in this thesis does the same.

Maternal effects

Maternal effects in SQG are the effects that the maternal progenitor has on the phenotype of their offspring that are unrelated to the offspring's own genotype (Danchin, 2011). These effects—caused by factors such as mitochondrial DNA transmission and *in utero* environment—can lead to the overestimation of heritability if the regression used is between offspring and maternal phenotypes. Since we define the maternal progenitor as a user's own early language there is clearly scope for such overestimation. Therefore, just as SQG regresses child onto father, we shall only compare Alice's main descendent with paternal progenitor Bob.

Assortative mating / homophily

Assortative mating is when progenitor mates are not randomly paired together, but systematically biased towards certain phenotypes (Falconer and Mackay, 1995). This is also known to occur within communication networks where it is known as *homophily*

—people are more likely to talk to those more similar to themselves in language phenotype, as well as to those with whom they have previously spoken (Bryden *et al.*, 2011; Danescu-Niculescu-Mizil *et al.*, 2011; McPherson *et al.*, 2001). Since assortative mating can inflate the similarity between generations when independent pairings are incorrectly assumed, this can also lead to overestimating heritability and must be controlled for in order to demonstrate language inheritance (Danescu-Niculescu-Mizil *et al.*, 2011). Adopting SQG control techniques of regressing *descendent trait minus maternal trait* onto paternal trait is therefore important.

Mimicry / imitation

Mimicry is the copying or replication of an observed behaviour and is often one of the first steps towards social learning. One suggested mechanism is the direct matching hypothesis—where areas of neurons are activated during an action, regardless of whether it is being observed or carried out. Evidence for this has been demonstrated in the premotor cortex of macaque monkeys (di Pellegrino *et al.*, 1992) and Broca's area —the motor area for speech—and other brain regions in humans (Iacoboni *et al.*, 1999).

As discussed above with regards to the psychological language literature, however, mimicry can be a purely reflexive action without any internalisation of the action. In order to demonstrate 'descent' with modification, the action must be observed in other contexts to show that the trait has truly been learnt. If a trait is observed in the same context that previous generation displayed the trait then mimicry cannot be excluded and descent cannot be demonstrated.

Although mimicry is not really a problem in SQG, it would analogous to measuring the heritability of thyroxine production by regressing the fetal concentration during the first trimester of pregnancy, before the fetal thyroid gland is fully functional (Korevaar *et al.*, 2016), against the maternal concentration at the same time. This extreme example displays problems beyond overestimation due to maternal factors. Because the fetus is actually using the maternal thyroxine, any estimate of heritability would be high even if thyrozine production were not at all heritable. It would be analogous to investigating the cultural heritability of washing behaviour in monkeys by assessing the student's behaviour during the lesson, rather than on a subsequent occasion when the teacher is

absent. It is because of this non-excluded mimicry falsely elevating heritability estimates that previous work has not provided sufficient evidence of language being heritable.

2.3.8. Language Summary

In summary, there is evidence that language contains culturally meaningful information and that it evolves at the population level (Pagel, 2009). There is also evidence at the individual level of vertical inheritance mechanisms (Nowak *et al.*, 2000) and population level evidence that a horizontal mechanism underlying evolution should exist (Eisenstein *et al.* 2012; 2014; Grieve *et al.*, 2016), as well as evidence of horizontal 'transmission' at the individual level (Bonin *et al.*, 2013; Brennan, 1996; De Looze, 2011, 2014; Hemphill and Otterbacher, 2012; Iwata and Wanatabe, 2013; Mitchell, 2012; Nenkova *et al.*, 2008; Pardo *et al.*, 2006, 2012; Stenchikova *et al.*, 2007; Ward, 2007, Włodarczak, 2013). There is, however, a lack of empirical evidence, at the individual level, of a horizontal inheritance mechanism—a mechanism important for a more comprehensive explanation of language evolution.

3. Methods

3.1 Overview

Although descriptions of the methods followed in this thesis—in sufficient detail for its reproduction—are outlined in the respective methods sections of each research chapter, the overarching methodologies that are repeatedly and consistently used are introduced in this methods discussion section. Furthermore, readers from any particular discipline may be unfamiliar with some the methods incorporated from fields outside their own. An overview and description of such methods, along with justifications for when and why applying them is useful, is also given in this methods discussion section of the thesis.

3.2. Overarching methodological design

The aim of this thesis is to tune in to the noisy, subjectively ambiguous and new terrorist signals provided on Twitter, social media and big data; using the advantages of big data to investigate patterns too subtle to have been amenable to prior study. As these signals are noisy, however, the primary concern of methodologies will be controlling for the noise generated by big data. Thus, in addition to the specific method logics that are introduced throughout this discussion of methods and the individual research chapter methodologies, some philosophical literature on falsifiability, reality and truth will be introduced to discuss how to check any patterns against appropriate controls.

The methodology used in this thesis is interdisciplinary, for three reasons. Firstly, scholars have investigated terrorism using methods from a range of disciplines, including: geopolitical, psychological (O'Connor, 2004; Rae, 2012; Silke, 2009), sociological (Huey and Witmer, 2016), computational (Berger and Morgan, 2015; Magdy *et al.*, 2015; 2016; Rowe and Saif, 2016) and historical (Laqueur, 2012; Veilleux-Lepage, 2016). This is both because the phenomenon of terrorism transcends these fields and the scarcity and patchiness of nearly all sources of terrorism data makes trying new approaches important. Thus, navigating the terrorism studies field, including placing this thesis amongst the existing literature, requires understanding and speaking the conceptual terminologies of a range of disciplines. Secondly, the nature of the dataset requires a variety of interdisciplinary approaches, in turn for several reasons; as although the large, complex and online social network data makes non-manual or

computational analysis the only viable approach, it is being interrogated for social patterns and behaviours with sociologically relevant questions and the contexts in which the data are generated and must therefore be interpreted, are driven by geopolitical factors. Finally, the members of the supervisory team come from departments of different academic disciplines, each with different terminologies and methodological approaches.

In spite of the desire to rigorously and objectively adapt and include methods from the social scientific, in addition to scientific, communities mentioned above, a fundamental aim of this thesis is to be quantitative and scientific in so far as is possible. Thus, the primary terminology in which it is written will perhaps be more familiar to data / computer scientists than to scholars of other disciplines.

3.3. Data and sampling

3.3.1. Twitter

The data used in this research has been sampled from the online social networking platform Twitter (http://twitter.com). Founded in 2006, Twitter primarily enables users to make public tweets of up to 140 characters, with options to join trending topics by including the—now infamous—'#' hashtag, or '@' mentioning other users. There are also features to follow the tweets made by other users and to send private messages. As of June 2016, Twitter has 313 million monthly active users (Twitter, 2016).

The social and linguistic phenomena investigated in this thesis, from language transmission to interaction between terrorists, are traditionally hard to observe. Studies translating a handful of spoken conversations, or contrived in laboratory settings devoid of social context, are rarely able to collect large quantities of data and are thus unlikely to detect the subtle patterns of language transmission.

On the other hand, widespread use of online social networks, such as Twitter, makes online conversations an excellent source of conversational language (Danescu-Niculescu-Mizil *et al.*, 2011). Furthermore, these free-to-download conversations, from across a wide range of topics, are objectively demarcated into discrete, time-stamped

text messages, thus avoiding the need for subjective transcription where the spelling and punctuation choices of the original author could be lost. While there are likely to be differences between online and offline conversations, the short, back-and-forth nature of online social media conversation streams are similar to offline spoken conversations. Part of the reason for this includes their relative immediacy, making them a better representation of the form and appearance of spoken language than some other commonly used sources of written language—for example emails or digitised books. They also more accurately reflect offline conversation with both the heterogeneity exhibited and the fact that the conversations cover a wider spectrum of interactions, "ranging from newly-introduced to old friends (or enemies)" (Danescu-Niculescu-Mizil *et al.*, 2011).

Although online conversations constitute only a fraction of human communication, they are an important model system for three key reasons. Firstly, online conversations are one of the few sources of language data large enough to be amenable to detailed quantification and statistical investigation. Secondly, irrespective of how much human communication or evolution of language is taking place though online conversations, we hypothesise that the laws of language transmission are fundamental enough to be visible even with this minor form of short, online transmission, given sufficiently big data. And finally, the aforementioned, important similarities with other conversations make them a good model for other conversations. It is this assumption of representativeness that allows us to generalise our findings, either language transmission patterns, or behavioural patterns observed within terrorist communication and extend them to other, more dominant, areas of communication that cannot themselves be easily studied.

Terrorism is a notoriously hard phenomenon on which to get data:

> *"one of the main problems in the study of the terrorist personality traits is that few terrorists wish to fill out a personality questionnaire"* (Webber, 2010).

Unlike governments with access to secret intelligence, academic researchers must often rely on secondary newspaper reports (Cooper, 1978; National Consortium for the Study of Terrorism and Responses to Terrorism (START), 2016; Silke, 2009) or the rare interview between a psychiatrist and those incarcerated having been convicted of terrorism (Rasch, 1979; Silke, 2009). Several long standing sources of intelligence,

including undercover human agents, covert bugging, interception of postal communications and surveillance and interception of mobile phone signals, emails and other digital internet communications, are exclusively available to governments and security service practitioners. HUMINT (human intelligence), ELINT (electronic intelligence), SIGINT (signals intelligence) and DIGINT (digital intelligence) (Aldrich, 2011; Andrew, 1990) are solely in the realm of intelligence agencies and law enforcement, inaccessible to academics working in a university or think tank setting.

Large volumes of representative data from online social networks, however, have made previously hard to study social phenomena amenable to investigation and this also appears to be the case within the apparently unique domain of terrorism; open access public messages on forums, social networks or video websites provide academics with OSINT (open-source intelligence) (Berger and Morgan, 2015; Brynielsson *et al.*, 2012; Lynch, 2014). The body of analysis that has grown around OSINT has led to at least two dedicated, annual symposiums: Open Source Intelligence & Web Mining (OSINT-WM) (since 2008; http://www.graphicslink.co.uk/IV2016/OSINT-WM.htm) and Foundations of Open Source Intelligence and Security Informatics (FOSINT-SI) (since 2012; http://fosint-si.cpsc.ucalgary.ca), although, as discussed earlier in the literature review, it was not a foregone conclusion at the outset of this work that terrorists should have been active on Twitter. Thus, there are a lot of Twitter signals available to researchers, investigators and experts—albeit weak and noisy signals. The aim of this research is, therefore, to tune in to the meaningful and predictive patterns amongst that social media noise. This thesis will investigate how best to analyse the traffic and tune in to terrorist signals.

As a side note on alternative online social media platforms that could provide a source of data: Facebook is similar, but less open, to research. The privacy settings are used more widely. Facebook and Google were also more aggressive in removing material as

> *"publicly traded companies with concerns about liability and a desire to create safe spaces for users"* (Stern and Berger, 2015, pp.134).

Whilst it is possible that different populations or subsets of the Baqiya family prefer to use different sites, or that they use a range of sites, each for a different purpose, that should not negate any Twitter findings, rather indicate that the generalisability and comprehensiveness of any conclusions may be limited.

3.3.2. Ethics

At the start of the research, in an initial risk assessment, it was decided that if, during the course of this research, any immediate or specific threat, or evidence of anything illegal, was discovered then this would be immediately reported to the appropriate police authorities. The data used in this thesis was collected using Twitter's standard, public API tools, not through any subversive means, or from any illegal sites. There were no consequences from any security services or the police. Even should surveillance services have registered this activity, there was nothing illegal about the work and we would have been happy to justify our work.

Regarding data protection, it was decided that beyond standard good practices, no special data protection rules applied to the raw Twitter data. Previous research has established that Twitter data is very clearly in the public domain. In addition, users signing up for the Twitter service agree to the terms and conditions, which include granting Twitter the right to "*make Content submitted to or through the Services available to other companies, organizations or individuals"* (Twitter, n.d.) via their API service (i.e. third party academics such as our research team). Furthermore, Twitter's terms and conditions clearly state that they do "*not disclose personally identifying information to third parties"* (Twitter, n.d.) and only share *"non-private"* information *"after we have removed any private personal information (such as your name or contact information)"* (Twitter, n.d.).

Since the raw Twitter data is in the public domain and special data protections apply to personal data where a individual "*can be identified from that information"*, no data protections were necessary beyond keeping it securely locked within the research office, on a password protected computer and only published in the form of agglomerated, anonymised summary results.

Where the experimental design required extremist content to be shown to individuals beyond the core research team, including undergraduates, specific ethical approval was sought and granted from the Royal Holloway University of London University Research Ethics Committee (REC ProjectID: 15). As part of this procedure, all data to be

presented was first examined manually and any particularly graphic content judged to have the potential to cause distress was excluded from further research.

Publishing findings on behavioural or linguistic trends exhibited by supporters of terrorism online, which by extension suggests how to identify the weak spots and possible counter-measures against terrorist groups, also raises ethical questions for us as academics. According to the Obama administration and the US Office of the Director of National Intelligence, when Osama Bin Laden was killed in the raid on Abbottabad in May 2011, he had with him multiple academic textbooks written by Western researchers (Office of the Directory of National Intelligence, 2015). Furthermore, as a direct result of the Snowden leaks, the Chief of MI6 stated that it was *"clear that our adversaries are rubbing their hands with glee, al-Qaeda is lapping it up"* (ITV Report, 2013), whilst the Director of GCHQ reported that al-Qaeda "*were moving to less vulnerable communication methods"* (ITV Report, 2013). Although one must remain critical of sources, this suggests that terrorist groups do take notice of academic research and that informing them of how they are vulnerable, even passively, is ethically questionable. Revealing information is not a new phenomenon, as was learnt when UK politicians read out the text of deciphered Russian cables during the first world war. The UK had learnt the lesson by the time they cracked ENIGMA in WWII that the advantages of disrupting one's enemy for a short period by releasing information is offset by the long-term challenge of a loss of continued intelligence. These concerns, however, are offset by an academic duty to publish openly and receive review, as well as an academic duty to contribute to holding the government to account by enabling public discussion with accurate information. The possibilities are to publish publicly, publish for government eyes only, or to leave the work to government and intelligence officials only. So long as each manuscript submitted for publication is considered carefully, both for the degree to which any findings can be useful to terrorists (or harmful to security), as well as what level of detail is published (for example not publishing datasets containing Twitter account handles or names), the risks associated with publishing the work in this thesis are low.

3.3.3. Python

Whilst a range of other methods to sample terrorism data from Twitter have been tried (Berger and Morgan, 2015; Huey and Witmer, 2016; Magdy *et al.*, 2015; 2016; O'Callaghan *et al.*, 2014), the interdisciplinarity of this research enables us to design the ideal datasets, identifying the appropriate features and questions needed to find them, but also to then build custom computer code to actually collect said ideal dataset, rather than having to rely on approximating, alternative, open-source tools (Huey and Witmer, 2016), or the spare time of external colleagues in the computer science department.

The computer code for sampling, downloading and saving data described above could be written in one of many programming languages. Throughout this thesis it is done using Python—as is much of the subsequent data analysis.

Python is one of the most intuitive, compatible and overall easiest to use programming languages for building bespoke data analysis scripts. Its simplicity makes it fast to write, it is interpreted (executes without the need for compiling) and its inbuilt memory handling makes it fast and efficient to run, so long as the datasets are no bigger in order than the tens of thousands. As a result of its open source nature, there are also a large number of free, well documented packages providing pre-built functions for a range of tasks and interactions with other software programs. Python is also very commonly used and thus help forums (e.g. https://stackoverflow.com) provide examples and explanations for nearly every sub-function, or at least those that have arisen during the course of this research. Building computational method script is therefore simplified to defining hypotheses, determining the necessary data to address them and then selecting the steps/functions needed to 'glue together' into a computational workflow.

Table 3.1. lists of some of the Python packages and libraries commonly used throughout this thesis that are crucial to being able to address the questions contained within. Table 3.2. gives an annotated example of a simple Python script to load some data and perform statistical tests on it.

Table 3.1.

Package	Description (link)
SciPy	"a Python-based ecosystem of open-source software for mathematics, science and engineering" (https://www.scipy.org)
NumPy	"the fundamental package for scientific computing with Python" (http://www.numpy.org)
matplotlib	"a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments" (http://matplotlib.org)
PyMongo	"a Python distribution containing tools for working with MongoDB and is the recommended way to work with MongoDB from Python" (https://api.mongodb.com/python/current/)
9.7 statistics	"This module provides functions for calculating mathematical statistics of numeric (Real-valued) data" (https://docs.python.org/3/library/statistics.html)
21.5 urllib, 21.6 urllib.request, 21.7 urllib.response	"a package that collects several modules for working with URLs" (https://docs.python.org/3.6/library/urllib.html)
Others	8.1 datetime, 9.2 math, 9.6 random, 16.1 os, 16.3 time, 19.2 json

Some of the Python packages and libraries commonly used throughout this thesis that are crucial to being able to address the questions contained within. Numbers are the Python package ID numbers.

Table 3.2.

```
#annotated (non-code) comments are indicated by the hash symbol
###IMPORT NECESSARY PACKAGES###
from pymongo import MongoClient
import numpy as np
from scipy import stats

def main():
      myconnection = MongoClient()         #connect to mongoDB
      my_db = myconnection.databaseName    #TO DO! change database
      my_collection = my_db.collectionName #TO DO! change collection

      data_x = [] #create an empty list to put the data in
      #the next line finds all records where 'field_1' equals '0'
      for record in my_collection.find('field_1' : '0'}):
            data.append(record['field_2']) #extract value of field_2

      data_y = [] #repeat the process for records with 'field_1' = '1'
      for record in my_collection.find('field_1' : '1'}):
            data.append(record['field_2'])

      doStatistics(data_x, data_y) #call the statistics function

def doStatistics(x,y): #calculates and prints means, stdevs and t-test
      print "mean x:", np.mean(x), "std. x:", np.std(x)
      print "mean y:", np.mean(y), "std. y:", np.std(y)
      t, p = stats.ttest_ind(x, y)
      print "2-tailed, unpaired, Student's t-test p-value:", p

###MAIN ENTRY POINT UPON RUNNING###
if __name__ == "__main__": #generic instruction to run main() function
    main()
```

A simple, example script. It first loads records from “collectionName” in the MongoDB database “databaseName”. It separately loads records with 'field_1' = '0' and '1'. It then calculates and prints the mean and std. for each set of data, as well as returning the p-value for a t-test of their difference.

3.3.4. Twitter APIs

The aforementioned custom computer code carrying out the sampling uses the range of Twitter APIs. APIs (Application Programming Interfaces) are ways of accessing online platforms (such as Twitter) as an alternative to manually viewing the site through a web browser. Each platform (that chooses to) builds their own, with their own set of standardised instructions, protocols and rules by which computers can automatically communicate with the platform and request, search for, upload and download data. These instructions are usually provided in the developers section of the platform's website and often provide specialised URL addresses for academics, businesses, application developers or other interested parties to build into their computer algorithms. In the case of Twitter's API, this is no different and each Twitter account can sign up for a free, unique set of bearer credentials with which their applications or algorithms can communicate directly with the Twitter servers. Twitter offers over 100 variations of their APIs, divided into 'REST' and 'Streaming' categories.

Streaming APIs

Streaming APIs (http://dev.twitter.com/streaming/) allow code or an application to open a channel (or 'stream') which remains open and connected to the Twitter servers. Twitter will then automatically send (or 'push') new data to the code whenever it becomes available—in as close to real time as possible. There are three types of Streaming API offered (Table 3.3.), depending on the objective, for example: Public stream—to get all new tweets on a given topic (e.g. "#tonightsElectionDebate); Public stream—to get all new tweets by another user (e.g. @BarackObama); User stream—to get all new tweets or private messages directed at the account who's credentials are being used to run the code; Site stream—to get all tweets directed at a range of accounts whose credential are all incorporated into the code or application. Within these streams, various filters can be used to define which tweets Twitter should send and which fields of metadata should be included or omitted: language of the tweet/user, user IDs, keywords, location of the users.

Table 3.3. The three types of Twitter Streaming API .

Public stream	"Streams of the public data flowing through Twitter. Suitable for following specific users or topics and data mining." (https://dev.twitter.com/streaming/overview)
User stream	"Single-user streams, containing roughly all of the data corresponding with a single user's view of Twitter." (https://dev.twitter.com/streaming/overview)
Site stream	"The multi-user version of user streams. Site streams are intended for servers which must connect to Twitter on behalf of many users." (https://dev.twitter.com/streaming/overview)

REST APIs

The Twitter REST APIs (http://dev.twitter.com/rest/public) are designed for any computational communication with Twitter that does not require real time data relayed to the researcher's computer as soon as it is posted to Twitter. Real time channels are more computationally expensive, using up more of Twitter's server and bandwidth, making dedicated REST APIs the preferred option wherever possible.

The REST APIs are further subdivided into a series of URLs (and associated request formats) that are dedicated to different types of Twitter information. For example, it is possible to use the 'GET users/show' REST API to request all of the information about one user's profile (or up to 100 at once with the 'GET users/lookup') whose unique Twitter IDs or handles (unique username beginning with '@') one includes in the request. An annotated example of a simple Python script using the 'GET users/lookup' REST API to request a user object is given in Table 3.4.

Table 3.4.

```
###IMPORT STATEMENTS###
import requests     #import the requests package
import json         #import the json package

#TO DO: this information needs to be inserted by the researcher
bearer_token = ""   #the unique Twitter API credential token
target_user_id = "" #id of the target-user whose information is wanted

#formats the authorisation credentials
header = {
      'Authorization': 'Bearer {}'.format(bearer_token),
      'Accept-Encoding': 'gzip',
      }

API_URL = "https://api.twitter.com/1.1/users/lookup.json?user_id="

#communicates with the Twitter API which responds to the request
API_response = requests.get(API_URL+target_user_id, headers = header)

#formats the block of text returned by Twitter into records and fields
data = API_response.json()
```

An annotated example of a simple python script to request a user object using the Twitter REST API.

It is also possible to closely replicate real time requests with the REST APIs, for example, to get all new tweets on a given topic (e.g. “#tonightsElectionDebate). To do this one could repeatedly use the 'GET search/tweets' REST API to return up to 100 tweets containing the keyword. By giving the unique Twitter ID of the oldest downloaded tweet and requesting only older tweets, it is possible to reconstruct a complete dataset of tweets that occurred between requests. This works because Twitter uses unique IDs for each object that increase monotonically with time. Although this approach requires more work in the code of the researcher, especially to ensure that no tweets are downloaded and saved twice should less than 100 tweets occur between requests, it places less effort on the Twitter servers. It is also the only possible approach to seek historical data (e.g. “#lastNightsElectionDebate”).

There are, however, limits on the rate with which Twitter permits requests for information. These rate limits differ by API type. Table 3.5. provides an indication, for several of the most relevant types of information, of the volume that can be downloaded in a given time window.

Table 3.5. Twitter REST API Rate Limits: Chart

Title	Requests / 15-min (user auth)	Requests / 15-min (app auth)	Use
GET users/lookup	180	60	Full user objects for up to 100 IDs.
GET search/tweets	180	450	Up to 100 tweets matching a query.
GET friends/ids	15	15	User IDs the specified user follows.
GET statuses/ user_timeline	180	300	200 most recent tweets by a user.

The rate limits for four relevant Twitter REST APIs. Information reproduced from http://dev.twitter.com/rest/public/rate-limits

Using the Twitter APIs to download large volumes of data is an alternative to the infeasible manual copy-and-paste approach, but also to algorithmic scraping of data from website source code. The latter, whilst the only option for platforms without their own APIs, requires detailed analysis of the layout of the html on the platform in question, sufficient proof that the desired information will always be in the same place in the scraped html and then the building of parsers and cleanliness checks. This is, in effect, doing all the work that goes behind building a custom API, but is far more susceptible to both programming errors and changes in the programming or layout of the platform.

Using Twitter's APIs directly (assuming one has knowledge of programming) is also more customisable than pre-built, open-source tools that usually utilise these APIs themselves anyway.

3.3.5. MongoDB

The data that is downloaded via the Twitter APIs is then saved and catalogued using MongoDB (http://www.mongodb.com). MongoDB is a free NoSQL (non relational) piece of database software. Data can be stored as 'records' within 'collections', themselves within 'databases'. Each record is assigned a unique '_id' number and a series of fields which can all be indexed for searching. Records can be inserted into a database collection, updated, deleted and searched; searches can be for one, more or all fields and can return the first, a random, or all identical matches.

Crucially, just as it is possible for direct communication between Twitter and computer algorithms via the Twitter APIs, MongoDB is easily compatible with direct computational communication via the MongoDB APIs. As the code written in this thesis is almost exclusively written in Python, the API used is PyMongo, however alternatives exist for other programming languages (including C, C#, C++, Java, Perl, Ruby, Hadoop and manual communication via the shell).

Data can thus be requested from Twitter via their APIs and then saved and catalogued directly into MongoDB for subsequent analysis. Once a computer has been given instructions on how to do this, it can do it repeatedly, automatically and continuously, thus generating a larger and better quality volume of data much more quickly and effectively than any human.

3.3.6. Snowball Sampling

There are a range of sampling strategies (as opposed to the range of computational methods by which to implement them) and the flexibility of writing our own sampling code gives us the freedom to choose between (or indeed to invent our own tailor made, custom solution).

One approach might be to use newspaper reports to construct an *a priori* list of accounts to sample. This approach, however, is biased to include only the most noteworthy 'official', 'media' or celebrity terrorist accounts that the media chooses to feature, a number limited to approximately 10-20 accounts. A second approach could be to use a list of accounts maintained by others (Berger, 2016; O'Callaghan, 2014), however this

depends on the reliability of the, often unpublished, methodology of the respective list authors. Although this approach is a potential last resort, if there is scope, time or expertise then tailor made sampling is preferable. A third, more standard approach, used by Berger and Morgan's census of ISIS on Twitter (Berger, 2015) as well as by Bryden *et al.* in other investigation of social media (Bryden, 2013; Tamburrini, 2015) is snowball sampling (Goodman, 1961).

Snowball sampling starts with a seed list of users and then samples all of the accounts at a distance of one (whether this distance is calculated by followers, friends, or messages) from any user in the seed list. The process is then repeated for all accounts at a distance of one from any account in the newly updated sample and so on.

Weighted snowball sampling biases sampling toward accounts with links to numerous other accounts that had already been sampled—for example, followed by more than 10% of the current sample (Wright, 2016). The reason for this is homophily: the tendency of people to associate with others similar to them (McPherson, 2001). This principle has been shown to lead to highly intra-linked communities on Twitter that bias their interactions to other members of the community and share a social identity (Bryden, 2011; 2013; Tamburrini, 2015).

3.4. Analysis

3.4.1. In defence of "algorithms"

The use of algorithms and computer scripts in this thesis are as nothing more than experimental and statistical workflows, designed in advance, that do not require human input at every single step. For computer scientists this may not seem like a problem, however, there is a body of social scientific and humanities literature expressing concern about the interpretability (Caruana *et al.*, 2015; Lipton, 2016), opacity (Burrell, 2016; Diakopoulos, 2014; Domingos, 2012; Gillespie, 2012; Seaver, 2014) and authority (Pasquale, 2009; Shirky, 2009) of algorithms.

The first concern is over *interpretability*. Some other forms of algorithms, such as neural networks or undirected machine learning algorithms, do partition, classify or

learn based on statistical measures such as partitioning of variance. The only evidence of their reliability, therefore, is their statistical reliability at predicting the teaching and test datasets. When it is not possible to *interpret* the processes the algorithm is following, or assign real-world constructs, variables or theory to the dimensions of variance, this lack of *interpretability* or *intelligibility* worries some authors (Caruana *et al.*, 2015; Lipton, 2016). With workflow algorithms of the type used in this thesis, however,—designed from first principles—this is not a problem.

Another area of concern is *opacity* (Burrell, 2016; Diakopoulos, 2014; Domingos, 2012; Gillespie, 2012; Seaver, 2014), where algorithms are neither published nor explained, often because they are commercial or governmental property. Indeed,

> "*opacity seems to be at the very heart of new concerns about 'algorithms' among legal scholars and social scientists*" (Burrell, 2016).

Again, this need not be a problem with this work, where each step in the methodology is clearly explained and justified and the algorithm merely carries out these steps.

Algorithms are neither some dark, magical form of "black art" (Domingos, 2012), nor are they mysterious "black boxes" (Lipton, 2016). Just as commonly used software such as SPSS enables one to compute textbook statistics, on volumes of data too large to analyse by hand, with as much *interpretability* and *transparency* as doing it by hand, a well designed algorithm is also an important tool for rigorous, non-manual analysis—albeit with more steps. Whilst the respective methods sections of the chapters explain 'what' was done rather the exact computational details of 'how' it was done, the scripts for calculating the Bray-Curtis index (String comparison metrics section), communicating with Twitter (Twitter APIs section) and loading data to test statistically (Python section) provide worked examples of how algorithms are used in this thesis.

3.4.2. String comparison metrics

Several of the questions in this thesis require a way to calculate how similar two users are and to monitor changes in similarity. As the source of data for this is their language, a language, or string, similarity measure is needed. There are several standard metrics to choose from (Gallagher, 1999; Gomaa, 2013), summarised in Table 3.6.

Table 3.6. String comparison metrics

Class	Examples	Notes
Character based	Longest Common Substring (LCS)	
	Damerau-Levenshtein distance[1]	[1]min. number operations to transform one string into the other
	Jaro distance[2]	[2]accounts for typical spelling deviations
	Jaro–Winkler distance[3]	[3]favours strings that match from the beginning for a set length
	Needleman-Wunsch algorithm[4]	[4]best alignment over the entire of two similar length sequences
	Smith-Waterman algorithm[5]	[5]local alignment over the conserved domain of dissimilar sequences with regions of similarity
	N-gram similarity algorithms[6]	[6]proportion of similar n-grams (sub-sequences of n items)
Term based (binary)	Jaccard distance[7]	[7] similarity a.k.a. Tanimoto similarity
	Sørensen–Dice similarity[8]	[8] proportion of shared words, a.k.a. Czekanowski's binary
	Ochiai coefficient[9]	[9]a.k.a. Cosine similarity
Term based (quantitative)	Bray-Curtis similarity[10]	[10] multiset proportion of shared words, a.k.a. Quantitative Sørensen–Dice/ Steinhaus/ Pielou's percentage/ Czekanowski's Quantitative
	Ruzicka similarity	
	Euclidean distance	
	Pearson's r	
	Block distance[11]	[11]a.k.a. Manhattan distance
	Kulczynkski's similarity	
Corpus based	Hyperspace Analogue to Language (HAL)	"Corpus-Based similarity is a semantic similarity measure that determines the similarity between words according to information
	Latent Semantic Analysis	
	Generalized Latent Semantic	

	Analysis Explicit Semantic Analysis Pointwise Mutual Information - Information Retrieval Normalized Google Distance DISCO	gained from large corpora" (Gomaa, 2013)
Knowledge based	Information content: Resnik, Lin, JCN Path length	"Knowledge-Based Similarity… using information derived from semantic networks" (Gomaa, 2013)

String comparison metrics divided into categories that indicate when they are best used. Compiled from (Gallagher, 1999; Gomaa, 2013).

The similarity metric adopted most in this work, however, is the Bray-Curtis index (Bray and Curtis, 1957). This measure, adapted from use comparing ecological habitats, is a suitable choice as it takes into account frequency, rather than simply binary presence/absence (Gallagher, 1999; Gomaa, 2013). Nor do tweets constitute a sufficiently large corpora to gain information to use corpus-based semantic similarity measures (Gomaa, 2013). Bray-Curtis similarity is calculated by first counting the number of words in both habitats or text samples, where multiple instances are counted multiple times. This count is then doubled (to represent both samples) and divided by the total number of words across the two samples. An annotated example script for calculating it is given below (Table 3.7.). To improve it, the words fed into the calculation are converted to lower-case and in some cases, stripped of punctuation. Despite these attempts to improve matching of words, it does not recognise cognates, synonyms or misspellings, although even a person's conscious or subconscious decision to use a different synonym, spelling or abbreviation of a word could potentially be carrying information worth building into the metric. Nor can the Bray-Curtis index distinguish between homographs (words that are spelled the same but have different meanings), which could misleadingly (but given a sufficiently large sample size probably not significantly) inflate the similarity between users. Despite these issues however, this rough metric, when combined with the volume of data, remains useful for demonstrating that phenomena exist, no matter how naïvely one searches for them. Part of the reason for this is that the majority of the information driving the metric comes from the grammar and structure inducing, common words (e.g. "and", "or", "but", "at",

etc.) (Bryden *et al*., 2013). Thus, changes in their usage that are normally invisible when other methods of analysis, such as interpretivist discourse analysis (see below) neither look for, nor detect, them.

Table 3.7.

```
def brayCurtis_Similarity(stringA, stringB):
      if (len(stringA) == 0) or (len(stringB) == 0):
            return float(0)

      lowercase_stringA = stringA.lower()   #convert to lower-case
      lowercase_stringB = stringB.lower()

      wordsA = lowercase_stringA.split(' ') #string to list of words
      wordsB = lowercase_stringB.split(' ')

      wordsA.sort()                         #alphabetically sort list
      wordsB.sort()

      A = len(wordsA)                       #count the words
      B = len(wordsB)

      a_counter = 0                         #initialise counters
      b_counter = 0
      overlap_score = 0

      #move through the two lists of words, comparing the word at the
      #head of each list. If they are the same, increase the score,
      #otherwise only move through the list that is alphabetically
      #behind.

      while (a_counter < A) and (b_counter < B):
            if wordsA[a_counter] == wordsB[b_counter]:
                  overlap_score += 1
                  a_counter += 1
                  b_counter += 1
            else:
                  if wordsA[a_counter] < wordsB[b_counter]:
                        a_counter += 1
                  else:
```

```
                    b_counter += 1

    return (float(overlap_score * 2) / float(A+B))

#print the output of running the above Bray-Curtis function on a
#series of test inputs
print  brayCurtis_Similarity("My name is","My name is")  #prints 1.0
print  brayCurtis_Similarity("My name is","my Name Is")  #prints 1.0
print  brayCurtis_Similarity("My name is","my name it")  #prints 0.667
print  brayCurtis_Similarity("My name is","my name")     #prints 0.8
```

A simple, example python script to calculate the Bray-Curtis similarity of two strings of text.

3.4.3. Discourse analysis

Discourse analysis (Fairclough, 2003; 2010; Schneider, 2013) is one of the most predominant methods in the social sciences and humanities fields. It involves the subjective, yet methodical, identification of themes, narratives and other patterns in texts or transcribed speech. These texts, interview transcripts, newspapers articles, film scripts, etc., are usually sampled from communities of similar individuals, or from texts discussing a shared topic, where the existence of a shared set of rules, vocabularies and narratives—i.e. a 'discourse'—is hypothesised. An important component of this type of analysis is the identification of not only 'what' is said, but what is 'meant' by it. This involves consideration of 'how' it is said, 'who' is saying it and the context in which they are writing or speaking.

Although, from a natural scientific viewpoint, the interpretivist nature of discourse analysis might seem subjective, when applied during this research, attempts have been made to adapt it in such a way as to make it quantifiable and comparable with other computational methods so as to be open to hypothesis testing and potential falsification (Popper, 1959).

3.4.4. Conformity with real world events

The purpose of this research is to find patterns and produce conclusions that are generalisable beyond the confines of Twitter. Whilst social interactions on Twitter are

sufficiently widespread to merit description as phenomena in their own right, it is as a proxy for offline, real world events that this research has the potential for the greatest impact.

Establishing the strength of the relationship between events on Twitter and events offline forms part of chapter seven, but this requires a source of data on real world events. The first possible source comes from newspapers and online media. Google News provides a centralised, comprehensive, free and reputable source of news articles compiled from the websites of newspapers around the world. As it is driven by Google, it is possible to do an advanced search of Google News for keywords and exact phrases to be included or excluded from an article's headline, body, URL or complete text, as well as to search within specific dates and by location and source. Despite this, since Google News also compiles articles from less reputable news sources, some criteria of credibility must be built into the workflow, such as grading the sources by reliability and then requiring a minimum number of each articles (with the threshold dependent on the quality of the sources of the articles) in order to conclude that there is sufficient cross-validated evidence to include a news story as fact. Using this method, it is also necessary to select the key search terms and this raises several questions: should the list be *a priori* or be allowed to grow organically if new words appear to be co-occurring with existing ones; how many synonyms should be included for each keyword; how many of the keywords should feature in an article in order for it to be included in the analysis; are some keywords more important in deciding this than others; should some, or all, of the articles be read manually to confirm that the presence of the keywords is in line with the expected topic of the article? Additionally, it is important to decide how many article to take, whether to take the first 100 for example, or everything containing the keywords,

An alternative method for identifying real world events is to use a list already maintained and constructed publicly; using public lists is an approach that has been used to monitor the Baqiya family (Berger and Perez, 2016; O'Callaghan *et al.*, 2014). Although this approach only works for topics where such a list exists, in those situations it simplifies the work needed to construct a training and test dataset, thus enabling more time to be concentrated on models to validate against the offline data. It also ensures

that the dataset can be publicly available for readers and other academic to satisfy themselves of its validity and/or use it for their own research.

The second approach is adopted in this this thesis for several key reasons. Firstly, the research questions relate to whether days during summer 2015 had events that were positive or negative for Daesh and thus there is good data already available via open source lists of events compiled from secondary news article sources, on Wikipedia (Table 3.8.). Consequently, given the methodological issues with the Google News approach, the fact that developing new models for analysing Twitter is the primary aim and the limited time remaining for the final chapter of the thesis, the best strategy was using the existing dataset and devoting the time to the Twitter models.

Table 3.8. Lists of Daesh related events on Wikipedia

Name of list	Events	References
Timeline of ISIL-related events (2013)	11	24
Timeline of ISIL-related events (2014)	216	489
Timeline of ISIL-related events (2015)	324	570
Timeline of ISIL-related events (2016)	163*	273*

The name of lists of Daesh related events, from 2013-2016, on Wikipedia, along with the number of events they contain and the number of secondary news article sources they cite. *as of 13th June 2016

3.4.5. Human annotation and 'truth'

Several of the questions addressed in this thesis involve the creation and testing of novel machine methods attempting to tune in to some important real world information via the proxy of Twitter data. Adopting standard machine-learning-evaluation approaches, all of these cases require data with which to train the machine and evaluate its performance (although in this work the data has trained a human who hypothesised potential machine models). Standard metrics are then used to compare the model's classifications against a test dataset of known, pre-characterised data. With this approach, although the pre-characterised dataset is assumed to represent the 'truth' and is a gold standard to be replicated (Cormack and Lynam, 2005; Yang and Srinivasan, 2014), fitted models are usually accepted to be only approximations—albeit useful approximations with predictive validity.

In studies of social networks, however, datasets are often ambiguous or subjective and data with which to rigorously evaluate performance. Methods of collecting a dataset of 'real world events' were discussed in the previous section, but both they and the datasets for characterising the 'real world people' behind Twitter accounts, suffer from problems of validity and truth outlined here. Currently, the standard approach to acquire pre-characterised training and test datasets is for humans to manually inspect and classify the data (Cormack and Lynam, 2005; Harris and Srinivasan, 2014; Yang and Srinivasan, 2014; Lo *et al.*, 2015). No matter how methodically or collaboratively done, the human annotated dataset against which models are fitted will also be an approximation. The fact that the "model and the data are two moving targets that we try to overlay one upon the other" (Rykiel, 1996) means that

> *"we cannot assume that data accurately represent the real system and therefore constitute the best test of the model"* (Rykiel, 1996).

In other domains, work has investigated how to improve human classification through crowdsourcing (Cormack and Lynam, 2005; Smucker *et al.*, 2012), machine methods (Lo *et al.*, 2015), or a mixture (Harris and Srinivasan, 2014). The Text Retrieval Conference (TREC) crowdsourcing track (Cormack and Lynam, 2005; Smucker *et al.*, 2012) is one example of such an effort. They invited attempts to develop crowdsourcing (later widened to include machine) methods to emulate human classification of documents. We adopt a similar approach, first evaluating the current gold standard of human classification and then developing and evaluating novel machine methods. As human annotation is currently the only method (and thus the gold standard) for annotating social media datasets, we argue that emulating it computationally is an important and worthwhile step forward.

3.4.6. Statistics

Most of the statistics used throughout this thesis are standard textbook tests. Normality testing is done with the Kolmogorov–Smirnov and depending on the result and the hypothesis being tested, the non-parametric Kruskal-Wallis/parametric one way ANOVA (analysis of variance), parametric Student's t-test/non-parametric Mann–Whitney U test, or parametric Pearson product-moment correlation coefficient/non-parametric Spearman's rank correlation coefficient.

To avoid the statistical issues that multiple hypothesis testing can cause (Bennett *et al.*, 2011) for data mining research, the approach taken in this thesis is to test specific *a priori* hypotheses. In cases where the hypotheses are tested multiple times, for example in multiple datasets or with multiple metrics, Bonferroni corrections (Bonferroni, 1936) are applied in order to prevent Type I errors.

There is another problem caused with analysis of large datasets such as those used in this thesis. Even small amounts of noise or difference occurring by chance, when combined with the volume of data, could be magnified to significance by standard statistical tests that make assumptions about the amount of variance or sample size. Therefore, an approach adopted in this work is to bootstrap random controls in order to establish the degree of noise that can occur by chance. Whereas testing for a significance of $p<0.05$ is conventional, i.e. to assume that there was a less than one in twenty chance that the result could have occurred by chance. By appropriately randomising the data, a negative control can be constructed without the condition being tested by the hypothesis. It is then possible to test whether the experimental data is more significant than the control data, as well as to quantify the amount of noise in the negative control data and by repeatedly bootstrapping it, calculate the probability of it being significant by chance.

4. How Humans Transmit Language

Overview in relation to the thesis

The first research chapter of this thesis addresses the validity of inferring conclusions from analysis of the language contained in tweets and Twitter profiles. Many forms of analysis involve studying the words, language and discourse expressed there; whether to characterise, infer hidden information, or monitor changes (Bryden *et al.*, 2013; Lo *et al.*, 2015). All of this work assumes that the language contained therein conveys phenomenologically or sociologically meaningful information and can thus act as a proxy for the offline user behind the text. In terrorism research in particular, when investigating changes in allegiances, sentiment or opinions, or evaluating changes to the landscape of groups, discourse or narratives, it is important to know whether any observed changes are connected to offline changes. The first step in demonstrating this is to show that interactions and behaviour are capable of influencing the language observed online. Chapter 4 investigates whether language can be horizontally transmitted and inherited; i.e. whether it is possible for people to change their language as a result of communicating with others; whether language can be influenced by and thus reveal, behaviour. I also show that part of a person's language is explained by their behaviour and interactions with others. Thus the assumption that analysis of online language allows us to tune in to some meaningful information about the users (particularly, to identify terrorism supporting users) is a valid one. This is different to previous work, much of which is done in the context of conversations between individuals who could simply be imitating one another (Hemphill and Otterbacher, 2012), suffers from small sample sizes, or both (Danescu-Niculescu-Mizil *et al.*, 2011).

Declaration of authorship

Chapter 4. How humans transmit language was co-authored with JB and VAAJ. The data for this chapter was downloaded by JB in 2009. The initial draft of the work was carried out by me. That initial analysis now constitutes the results in Fig. 4. and the initial literature review informs both the current draft and the language section in the literature review at the start of this thesis. After subsequent discussion between JB, VAAJ and myself, *VAAJ devised the model in the supplementary material* of Chapter 4 and *JB carried out the modelling analysis (Fig. 3.)*. The analysis forming Fig. 2. was initially conceived and carried out by me on a sample of around 100 words and *subsequently repeated by JB on the 1,000 words in the manuscript*. Although I wrote the first draft and reviewed the literature, *JB re-started, formatted, produced figures for and led the writing of, the current draft*. All authors, including myself, edited and critiqued the manuscript. Overall, all authors, including myself, contributed equally to this chapter.

Bryden, J., Wright, S., Jansen, V.A.A., 2016. How humans transmit language. [*Submitted*] 1-5

How humans transmit language

John Bryden[1], Shaun P. Wright[1], Vincent A. A. Jansen[1]

[1]*School of Biological Sciences, Royal Holloway, University of London, Egham, TW20 0EX, United Kingdom*

Language transmission plays a key role in linguistic and cultural evolution[1–11]. As evolution requires inheritance we need to identify what exactly is transmitted. This means we must go beyond short-term studies of language change within a dialogue[12–16], which can be transient or reflective, and build a mechanistic understanding from empirical evidence of lasting changes. Here we use data from natural conversations to uncover and validate a process of language transmission whereby individuals change their word-usage frequencies upon conversations with their peers and then use the updated frequencies when talking to different conversation partners. The data are consistent with the existence of an internal representation of word frequencies. This suggests a simple mechanism of language transmission: people hold variable affinities for words, and update these affinities based on the language they experience. We note that the process of storing and passing on of information during conversation, which we have presented here, is analogous to that of genes being passed on and can be analysed with techniques from genetics. Our results link language transmission at the individual level to changes at the level of linguistic human communities, and can

help placing such changes in time. This provides a solid basis to understand and quantify language transmission and evolution.

Language use is constantly in flux and changes in language use happen at many spatial and temporal scales. Historical evidence shows how population groups experience wholesale changes in word usage and language syntax across many generations[1–4,17]. Based on this, a broad theoretical background has been developed which explains how these large-scale and dynamic language patterns can be generated by language change at the individual level[1–11]. These studies assume that language elements are repeatedly transmitted between individuals in a population, and then use mathematical models or computer simulations to show that a macroscopic language pattern is generated from iterations of this individual behaviour. This makes it plausible that macroscopic changes follow from an accumulation of individual transmission events. However, to move beyond such 'plausibility arguments'[8] we need a mechanistic understanding of how language is transmitted.

At an individual level, we adopt elements of our language throughout our lives. As children we acquire the majority of our language from our parents, but as we grow older we increasingly pick up language from our peers[17]. This form of cultural transmission between peers is called horizontal transmission[18]. While language acquisition early in life (known as vertical transmission) can be easily observed, the effect of horizontal transmission later on is more subtle and more difficult to detect. It has been known for several decades that word-usage patterns, as well as other linguistic variables, are imitated between interlocutors[12–16,19].

This imitation can be transient or reflective. To look for lasting changes we need to look for iterated transmission where people adopt words and use them in other conversations, which has been observed under laboratory conditions[10]. How language elements transmit in a lasting way between peers in natural situations is hard to measure, in part because there is a weak effect per conversation.

Here, we will provide evidence of horizontal language element transmission using a large corpus of online conversations. This 'big data' approach allows us to investigate how word frequencies are horizontally transmitted in a lasting way in uncontrolled communication within a population, and detect and quantify this process despite the weakness of the signal. The transmission of language elements is often assumed to be analogous to the spread of genetic traits[4]. We therefore use techniques from the toolbox developed within evolutionary biology, on the interface between population genetics and linguistics[18,20]. We study horizontal language transmission by investigating the change in the use of words following exposure to the language of other people. This assumes that, beyond simply having a lexicon, we have some internal language representation which influences which words we choose and how often we use them[20]. We cannot directly observe this representation, but we can infer it from word usage frequencies in a person's outgoing communication. We will show here how it is possible to identify a change in the representation over time and then show, using advanced statistical methods, that this change happens due to conversations with another individual.

We will use a simple model for the internal representation of language which incorporates

transmission of language between individuals. Because our aim is to study how word frequencies change, this highly simplified internal representation does not place any specific importance on grammar, syntax or word order. We simply treat communication as a multiset or a 'bag of words'[21]: how often a person uses a word is reflected by the number of copies of the word in their bag. Word instances received from conversation partners can occasionally replace other words in the bag, changing the internal representation and allowing the frequency of stored words to change in response to conversation (see Fig. 1). This model forms a Moran process and can be analysed using well understood techniques[22]. Our analysis of the model (see Supplementary Information) shows how the word frequencies used will equilibrate over time towards the frequencies received from conversation partners in a way that is very similar to osmosis (Fig. 1). The model predicts that an individual's word-usage patterns change through conversations with others and that this change will manifest itself in the word frequencies that the individual then uses to other people. Although in this model language changes in response to all language received, the effect of a conversation with a particular conversation partner will leave its mark, even if this conversation is only a relatively small part of all their conversations.

Results

We first show that word frequencies used by an individual change in response to the language used by a conversation partner, as predicted by our model. We studied a data set of conversations formed from a sample of 200 million messages sent publicly between users of

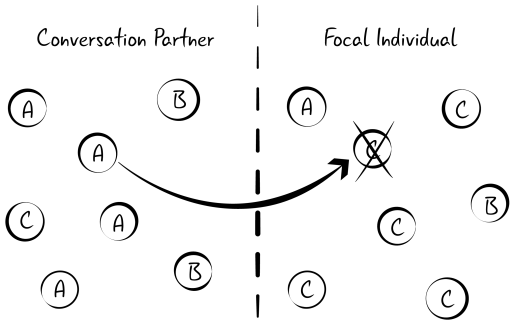


Figure 1: **An osmosis-like process for horizontal language transmission used in our model.** The two halves of the diagram show the internal language representations of two individuals as containers. The figure shows how an individual in our framework copies and stores a word from their conversation partner; an instance of word A is incorporated, replacing an instance of word C. The number of instances of a particular word defines how likely someone is to use the word in a given situation. Word frequencies change similarly to osmosis in that over time the frequencies of words in both halves will tend to equilibrate.

the Twitter web site[19] (see Methods). To eliminate any transient imitation that others have found in online communication[15,16], we excluded any mutually directed messages between a pair being studied in our analysis. Motivated by the result from our model that the difference between users is important, we looked at the influence that the difference between a focal user and their partner's early usage of a word has on any later change of the focal user's usage of the word. Since this is mathematically related to the heritability of genetic traits[23], controlled for assortative mating and regression to the mean, we dub this *word-heritability*. Over the 1,000 words tested (see Methods), we found that mean word-heritability was significantly greater for pairs of users that had sent each other messages than for control-pairs that had not (see Fig. 2). This indicates that an individual changes their word-usage toward that used by their conversation partner.

Within our model, when a focal individual encounters word instances used by another individual, some of these incoming word instances will be incorporated replacing word instances within the focal individual's internal representation. The *incorporation rate* (α) can be measured. To do so, we implemented the model as a stochastic process where we maintain an individual's representation of a word over time, calibrated by the individual's usage of the word. We test whether incorporating incoming words from a conversation partner will increase our ability to predict future out-going usage of the focal individual by calculating the likelihood of the focal individual's usage of the word based on our maintained internal representation (see Supplementary Information for precise details). We tested 1,000 different words (see Methods) and found the most likely value of α for each word. We then tested if the

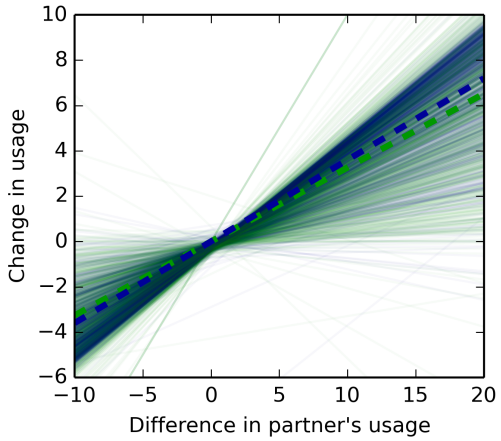


Figure 2: **Word heritability between conversing partners is greater than that for non-conversing partners.** For each test word, we plot regressions (see methods) for data from conversing partners (blue lines) and non-conversing partners (green lines). The regression lines were superimposed by translucently plotting lines for each regression, interleaving between the two data sets. The two dashed lines are the mean slopes regressed over data collected for all of the words. The difference between the mean slopes, a measurement of word-heritability, was 0.0340 ($p < 9.5 \times 10^{-10}$, Mann-Whitney U test). This analysis was done on individual words; we also checked the result with a regression across all the sample words, using a bootstrap (see methods) to test that the mean slope was greater for conversing partners than for non-conversing partners ($p < 0.001$).

incorporation process is dependent on word frequency[24]: i.e., the per-instance incorporation rate for a given word depends on the population's usage frequency of the word. Interestingly, we found that the rate of a word instance being taken up in our model is independent of word frequency across a wide range of word frequencies (see Fig. 3). This indicates that we are as likely to adopt an instance of a frequent word as much as we are to adopt an instance of an infrequent (and therefore conversation specific) word. This suggests that we have found a perspective whereby word transmission is a neutral process; a view consistent with the heavy tailed distributions of word frequencies predicted by Zipf's law[22].

Our model also predicts that the more two users communicate with each other, the less different their overall word usage frequencies will be. We tested this prediction by looking at whether two users increasingly shared more of their words over time if they sent one another higher numbers of messages. We found a highly significant, positive correlation between the change in the proportion of word instances shared between two users and the number of messages sent between them; as well as a close quantitative fit with our model (see Methods) and the data (Fig. 4). The value of the word incorporation rate, α, found was 0.01, a similar order of magnitude to the mean incorporation rate found in Fig. 3. These measurements indicate that we subconsciously incorporate approximately one in every 100-200 words that we experience.

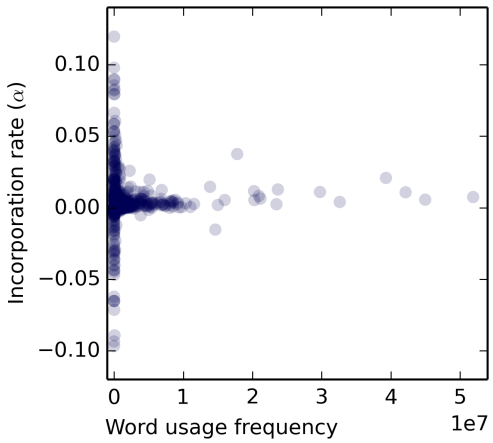


Figure 3: **The rates with which words are incorporated is independent of usage frequency.** Each circle is a word's incorporation rate (circles have translucency of 30%). Linear regression finds no correlation between the word's usage frequency (over the whole sample) and the incorporation rate (two-tailed Pearson's: r^2=0.00040, p=0.54). The mean value of the word incorporation rate α is 0.0043, which we found significantly greater than zero (p=0.0083, bootstrapping with 10,000 resamples of 100 values, and calculating the proportion of resamples with mean greater than zero). The high variance for very low frequencies is due to sampling effects.

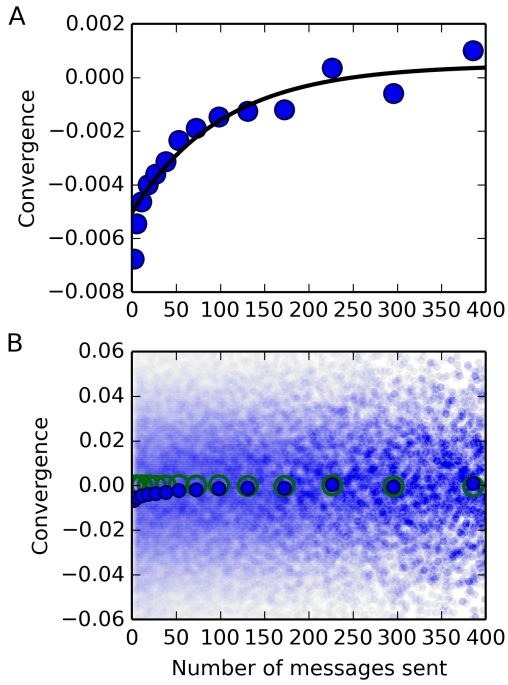


Figure 4: **The more messages were sent between two users, the more their language converged.** Panel **A** plots the means of bins of conversation pairs (binned along the x-axis showing x, y means of each bin) and the fitted model prediction (black line, see Methods). The line crosses zero at approximately 310 messages sent. Panel **B** illustrates the large variance in the data (unbordered translucent circles which are superimposed). The convergence of 500 conversation pairs (sampled with replacement) are plotted per bin on the x-axis (bordered blue circles). Control values are also shown (bordered green circles).

Discussion

Our results demonstrate that humans adopt lasting changes in their language usage upon conversation. These changes are consistent with the existence of an internal representation of word frequencies, where words are incorporated in a Moran process. We found that the rate at which an instance of a word is incorporated is independent of how frequently the word is used. Put together, this means that the more two individuals converse, the more they will use similar language outside their conversations. A corollary of this is that the word usage of two isolated, or weakly connected groups, will drift apart on this time scale.

The use of large quantities of data, gleaned from online conversations, allows us to detect evidence for an underlying process of language transmission. Through identifying this process, we fill a gap in our understanding of how language is shaped and evolves[4,9,25,26]. We demonstrate a process which has subtle effects at the individual level (see Fig. 4b). However, when this process is iterated many times within a population, large scale social patterns can develop. For instance, it follows from our results that groups which interact more with one another will share similar and distinctive language patterns; which is borne out by evidence from online conversations[19]. Furthermore, the ability to track the changing language usage of groups as they become more or less isolated can yield tools for dating changes in language use, inferring changes in population structure and even predicting such changes in the future[2].

The process of transmission demonstrated here, being peer-to-peer in nature, forms a basis for horizontal transmission[18]. Indeed, our results reject a model that human language use

can solely be explained by vertical transmission as we have shown that horizontal transmission does take place. Furthermore, the mechanism of lasting transmission we have identified can go beyond horizontal transmission and may underlie vertical transmission whereby children acquire vocabularies from their parents. From this perspective, we propose that vertical transmission can work in much the same way as horizontal transmission but with an inequality between parents and children whereby parents are much less likely to pick up words from their children than *vice versa*. With an understanding of both forms of transmission, the model and evidence that we have presented can be applied to understand how word frequencies can change across several generations of a population.

Language transmission is a cognitive process with an underlying neurological mechanism. Our evidence that word frequencies are transmitted from person to person points to insights which can inform neuroscience about the sorts of brain structures, mechanisms and memory that are necessary for language uptake and storage, and may be awaiting discovery. For example, an internal, mutable representation of word frequency suggests a reinforcement process and directs neuroscientists towards plasticity theories; a conclusion supported by various studies showing a role for plasticity and/or Hebbian learning in language therapy[27], acquisition[28] and processing[29,30].

There are no genes for words, or other specific language features, yet languages change in a way that is very reminiscent of biological evolution. This suggests that there is something which is inherited and which is passed on like a gene, even if we do not what this something is.

Here we show how word frequencies can be stored and passed on. This forms a quantifiable basis for studying descent with modification of language: a requirement for language evolution.

Methods

Data acquisition. We used conversations between users recorded on the social networking site Twitter. Online conversations on social networks allow the observation of natural, everyday language within its social context in a way that more formal, written media does not. The informal style of this language, and its short, back-and-forth nature, makes it much closer in form and appearance to spoken language than most other forms of written language. Communication on Twitter replicates the heterogeneity in usage that is found in spoken language[11,16,19]. The ubiquity of the use of online social media for human interaction allows the gathering of this data at a large scale and in quantities that are not normally achieved for spoken language. While there are likely to be differences, Twitter conversations are more like regular conversations than other, written forms of communication.

The data were recorded from the Twitter website during December 2009. A snowball sampling process was used to gather users as follows: for each user sampled, all their tweets that mentioned other users (using the '@' symbol) were recorded and any newly referenced users were added to a list of users from which the next user to be sampled was picked. Starting from a random user, conversational tweets (time-stamped between January 2007 and November 2009) were sampled, yielding over 200 million messages from over 189,000 users. We ignored messages that were copies of other messages (so called retweets, which are

identified by a case-insensitive search for the text 'RT').

Test words. The following tests were done using a list of 1,000 different test words. These words were selected randomly from the complete collection of all text in the sample.

Word heritability analysis. Messages were temporally split into 'early' and 'late' halves around the median time. An 'early sample' was created by randomly sampling 1,000 words from the amalgamated early tweets. This was repeated with the amalgamated late tweets to create a 'late sample'.

Word heritability was measured by regressing over a series of points: each calculated on the basis of a single given word, and a randomly shuffled pair of users. For the first axis of the regression, we recorded the difference of the first user's usage of the word compared with that of the other user during the two early halves. For the second axis, we recorded the amount which the first user changed their usage of that word over time between their early and late halves. Two regressions were plotted for each word: one for conversing partners and one for non-conversing partners. In all we recorded approximately 500 million data points between conversing partners and 9 million data points between non-conversing partners.

To test for significance, we did a bootstrap by generating two resamples of 500K points from the conversing and non-conversing data sets and regressed a line through each sample. We then measured the difference between the two slopes and recorded the proportion (reported in the main text as p) of the 1,000 bootstrap resamples for which the slope for non-conversing individuals exceeded the slope for conversing individuals. As a test, we confirmed that similar

results could be achieved with smaller resample sizes.

Convergence analysis. Convergence between a pair of users was measured by calculating the Bray-Curtis similarity[31] of the pair's late samples, and then subtracting the Bray-Curtis similarity of their early samples.

The model fitted to the data was Eq. (1) from the Supplementary Information:

$$y = c_1 + c_2 e^{-\alpha x}$$

Fitting was done against the points sampled for display in Fig. 4 using a least squares method. The values found were: $c_1 = 0.000478, c_2 = -0.00552, \alpha = 0.00982$.

References

1. Dunn, M., Terrill, A., Reesink, G., Foley, R. A. & Levinson, S. C. Structural phylogenetics and the reconstruction of ancient language history. *Science* **309,** 2072–2075 (2005).

2. Lieberman, E., Michel, J.-B., Jackson, J., Tang, T. & Nowak, M. A. Quantifying the evolutionary dynamics of language. *Nature* **449,** 713–716 (2007).

3. Gray, R. D., Drummond, A. J. & Greenhill, S. J. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* **323,** 479–483 (2009).

4. Pagel, M. Human language as a culturally transmitted replicator. *Nat Rev Genet* **10,** 405–415. ISSN: 1471-0056 (June 2009).

5. Nowak, M. A., Komarova, N. L. & Niyogi, P. Evolution of universal grammar. *Science* **291,** 114–118. ISSN: 0036-8075 (Jan. 2001).

6. Nowak, M. A., Komarova, N. L. & Niyogi, P. Computational and evolutionary aspects of language. *Nature* **417,** 611–617. ISSN: 0028-0836 (June 2002).

7. Steels, L. & Kaplan, F. Aibo's first words: The social learning of language and meaning. *Evolution of communications* **4,** 3–32 (2002).

8. Castellano, C., Fortunato, S. & Loreto, V. Statistical physics of social dynamics. *Rev Mod Phys* **81,** 591–646 (May 2009).

9. Chater, N. & Christiansen, M. H. Language Acquisition Meets Language Evolution. *Cogn Sci* **34,** 1131–1157. ISSN: 1551-6709 (Sept. 2010).

10. Kirby, S., Griffiths, T. & Smith, K. Iterated learning and the evolution of language. *Curr Opin Neurobiol* **28,** 108–114. ISSN: 0959-4388 (Oct. 2014).

11. Eisenstein, J., O'Connor, B., Smith, N. A. & Xing, E. P. Diffusion of Lexical Change in Social Media. *PLoS ONE* **9,** e113114. ISSN: 1932-6203 (Nov. 2014).

12. Brennan, S. E. Lexical entrainment in spontaneous dialog. *Proceedings of ISSD* **96,** 41–44 (1996).

13. Pickering, M. J. & Garrod, S. Toward a mechanistic psychology of dialogue. *Behav brain sci* **27,** 169–190 (2004).

14. Gallois, C., Ogay, T. & Giles, H. in *Theorizing About Intercultural Communication.* (ed Gudykunst, W. B.) 121–148 (Sage, Thousand Oaks, CA, 2005). <http://espace.library.uq.edu.au/view/UQ:72030>.

15. Danescu-Niculescu-Mizil, C., Gamon, M. & Dumais, S. *Mark My Words!: Linguistic Style Accommodation in Social Media* in *Proceedings of the 20th International Conference on World Wide Web* (ACM, New York, NY, USA, 2011), 745–754. ISBN: 978-1-4503-0632-4. doi:10.1145/1963405.1963509. <http://doi.acm.org/10.1145/1963405.1963509>.

16. Tamburrini, N., Cinnirella, M., Jansen, V. A. A. & Bryden, J. Twitter users change word usage according to conversation-partner social identity. *Soc Networks* **40,** 84–89 (2015).

17. Bloomfield, L. *Language* ISBN: 978-0-226-06067-5 (University of Chicago Press, 1933).

18. Cavalli-Sforza, L. L. & Feldman, M. W. *Cultural transmission and evolution: a quantitative approach* **16.** <https://books.google.co.uk/books?hl=en&lr=&id=pBGvyNXkWYcC&oi=fnd&pg=PR5&dq=cultural+transmission+and+evolution+a+quantitative+approach&ots=3LzZIgC6p_&sig=qe3V69pN7be60D7syXtzzQxVrwM> (Princeton University Press, 1981).

19. Bryden, J., Funk, S. & Jansen, V. A. A. Word usage mirrors community structure in the online social network Twitter. *EPJ Data Science* **2,** 3 (2013).

20. Wang, W. S.-Y. Language Change. *Ann N Y Acad Sci* **280,** 61–72. ISSN: 1749-6632 (Oct. 1976).

21. Salton, G. & McGill, M. J. *Introduction to modern information retrieval* ISBN: 978-0-07-054484-0 (McGraw-Hill, 1983).

22. Blythe, R. A. Neutral evolution: A null model for language dynamics. *Advances in Complex Systems* **15,** 1150015. ISSN: 0219-5259, 1793-6802 (May 2012).

23. Falconer, D. S. & Mackay, T. F. C. *Introduction to Quantitative Genetics* 4 edition. ISBN: 978-0-582-24302-6 (Longman, Dec. 1995).

24. Church, K. W. *Empirical Estimates Of Adaptation: The Chance Of Two Noriegas Is Closer To P/2 Than P2* in *Proceedings of the 18th conference on Computational linguistics-Volume 1* (Association for Computational Linguistics, 2000), 180–186. <`http://dl.acm.org/citation.cfm?id=990847`>.

25. Croft, W. *Explaining Language Change : an Evolutionary Approach* ISBN: 978-0-582-35677-1 (Pearson Education, Jan. 2000).

26. Pagel, M., Atkinson, Q. D. & Meade, A. Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature* **449,** 717–720. ISSN: 0028-0836 (Oct. 2007).

27. Sarasso, S. *et al.* Plastic Changes Following Imitation-Based Speech and Language Therapy for Aphasia A High-Density Sleep EEG Study. *Neurorehabil Neural Repair* **28,** 129–138 (2014).

28. Kim, K. H. S., Relkin, N. R., Lee, K.-M. & Hirsch, J. Distinct cortical areas associated with native and second languages. *Nature* **388,** 171–174. ISSN: 0028-0836 (July 1997).

29. Chee, M. W., Hon, N. H., Caplan, D., Lee, H. L. & Goh, J. Frequency of concrete words modulates prefrontal activation during semantic judgments. *Neuroimage* **16,** 259–268 (2002).

30. Wennekers, T., Garagnani, M. & Pulvermueller, F. Language models based on Hebbian cell assemblies. *J Physiol Paris* **100,** 16–30. ISSN: 0928-4257 (Sept. 2006).

31. Bray, J. R. & Curtis, J. T. An ordination of the upland forest communities of southern Wisconsin. *Ecol Monogr* **27,** 325–349 (1957).

Acknowledgements The authors contributed equally to this article.

Competing Interests The authors declare that they have no competing financial interests.

Correspondence Correspondence and requests for materials should be addressed to John Bryden (john.bryden@rhul.ac.uk) or Vincent Jansen (vincent.jansen@rhul.ac.uk).

HOW HUMANS TRANSMIT LANGUAGE: A MODEL FOR LANGUAGE INTERNALISATION AND TRANSMISSION.

J. BRYDEN, S. P. WRIGHT, AND V. A. A. JANSEN

1. The bag of words model

We will present a basic model for language transmission. The basic premise is that language production is based, in part, on an internalised distribution of words. This internalised distribution will be formed dependent on the frequency of words in the language that an individual is exposed to, and this internalised distribution will in turn be used for the production of language. We will assume that this part of the internalised language functions like a bag (also known as a multiset): a set of which can have multiple instances of its elements can appear more than once. In a bag of words the same word can appear multiple times. We assume that the bag has got a fixed size.

Language is produced by sampling from this bag. As a consequence, for sufficiently long text the word frequency in the text will tend to the frequency in the bag. Language is internalised through sampling from the words that an individual receives. Our model assumes that words are sampled from the incoming text, and placed in the bag. To keep the size of the bag constant, a random element is then removed from the bag.

Individuals are in contact, and communicate with other individuals. These contacts span up a network on which communication takes place and messages are exchanged. The words produced are sampled from the bag of words of the individual that sends the message. A sample of these words will be internalised by the receiver of the message.

We consider a population of n individuals. Individual i sends words to individual j with rate r_{ij}. Let the bag of words of individual i be given by the set $\{x_{i1}, \ldots, x_{iw}\}$ where the x_{ik} is number of copies of the kth word that individual i has in its bag of words. The numbering of the words is the same for all individuals and the w is the total number of words in the population. The size of the bag, s, is constant and the same for all individuals, and hence $\sum_{k=1}^{w} x_{ik} = s$ for all $1 \leq i \leq n$.

Through the internalisation of words received, the internalised word numbers $x_{k,i}$ change over time. We will describe this change through a stochastic process similar to a Moran process (Blythe 2012, BLythe and McKane 2007). The Moran process, in its simplest form, describes the change in a population of genes through selection and random drift caused by replacement through birth in a finite population. Here, we will apply a similar logic to describe the changes in the bag of words. Words are internalised with a rate proportional to the words received. Individual i receives messages from individual j at rate r_{ij}, and we will assume that individuals do not act on messages received from self and hence take all $r_{ii} = 0$. The rate with word m is internalised is $\alpha_m \sum_{j=1}^{n} r_{ij} \frac{x_{jm}}{s}$, where α_m is the rate constant for word m, or the

probability per word of type m received that it will lead to an internalisation event. In every internalisation event a random word is removed from the receiver's bag. The rate with which a copy of word k is removed is therefore $\alpha_m \sum_{j=1}^{n} r_{ij} \frac{x_{jm}}{s} \frac{x_{ik}}{s}$.

Thus the event:

$$x_{ik} \to x_{ik} - 1; x_{im} \to x_{im} + 1$$

takes place with rate

$$\alpha_m \sum_{j=1}^{n} r_{ij} \frac{x_{jm}}{s} \frac{x_{ik}}{s}$$

1.1. **Simulating change.** The model above is a generic description of the transmission and internalisation of word frequencies. It can be used to simulate networks of communicating individuals to study their changes in word frequencies. This can be done as follows: first, set up a network through specifying the rates of communication r_{ij}. Initialise the network by assigning an initial bag of words to all nodes on the network, which specifies the network at $t = 0$. Internalisation events take place with total rate

$$\sum_{i=1}^{n} \sum_{m=1}^{w} \sum_{k=1}^{w} \alpha_m \frac{x_{ik}}{s} \sum_{j=1}^{n} r_{ij} \frac{x_{jm}}{s} = \sum_{j=1}^{n} \sum_{i=1}^{n} r_{ij} \sum_{m=1}^{w} \alpha_m \frac{x_{jm}}{s}.$$

Using the Gillespie algorithm, the waiting time until the next event is exponentially distributed with mean

$$\left(\sum_{j=1}^{n} \sum_{i=1}^{n} r_{ij} \sum_{m=1}^{w} \alpha_m \frac{x_{jm}}{s} \right)^{-1}.$$

We can draw the waiting time until the next event and update time from this distribution. The chance that the next event is in individual i is

$$\left(\sum_{j=1}^{n} r_{ij} \sum_{m=1}^{w} \alpha_m \frac{x_{jm}}{s} \right) \left(\sum_{j=1}^{n} \sum_{i=1}^{n} r_{ij} \sum_{m=1}^{w} \alpha_m \frac{x_{jm}}{s} \right)^{-1},$$

and use this to identify i. Word k is removed with probability $\frac{x_{ik}}{s}$, use this to lower the number of x_{ik} by one. Refill the gap with word m. Word m is now chosen with probability

$$\left(\alpha_m \sum_{j=1}^{n} r_{ij} \frac{x_{jm}}{s} \right) \left(\sum_{j=1}^{n} r_{ij} \sum_{m=1}^{w} \alpha_m \frac{x_{jm}}{s} \right)^{-1}.$$

Note that these expressions simplify considerably if the internalisation rates are the same for all words: if $\alpha_m = \alpha$ then $\sum_{m=1}^{w} \alpha_m \frac{x_{jm}}{s} = \alpha$

1.2. **Ensemble means of word frequencies.** We will next derive the behaviour of the system if it would be averaged over many simulations, of the same network, all starting from the same initial conditions. We will refer to one such simulation as a single realisation of the process. To find the average over many realisations we will write down the master equation for this process.

To do so we will consider the probability of the system to be in a certain state as the process develops. The system consist of n individuals, who each have bags of words with s spaces, and we need to keep track of all possible permutations

of words possible. To help us do the book keeping we need some extra notation. All communication rates can be put together in a $n \times n$ matrix $\mathbf{R} = (r_{ij})$. We will represent the state of the system by a $n \times w$ matrix $\mathbf{X} = (x_{im})$, which has as elements the word counts of all individuals. Let $P(\mathbf{X})(t)$ be the probability for the system to be in state $\mathbf{X}$ at time t. This probability changes over time according to the master equation, given by:

$$\frac{\mathrm{d}P(\mathbf{X})}{\mathrm{d}t} = \sum_{i=1}^{n}\sum_{j=1}^{n} r_{ij} \sum_{m=1}^{w} \alpha_m \frac{x_{jm}}{s} \left(\sum_{k=1}^{w} \frac{x_{ik}+1}{s} P(\mathbf{X} - \mathbf{E}_{i,m} + \mathbf{E}_{i,k}) - \sum_{k=1}^{w} \frac{x_{ik}}{s} P(\mathbf{X}) \right)$$

where $\mathbf{E}_{i,m}$ is a $n \times w$ matrix, in which all elements are zero, expect from element i, m, which is one. We will use the master equation to calculate how the ensemble mean of the word frequencies change over time. The ensemble mean of all the word counts of all individuals $\mathbf{X}$ is given by $\sum_{\mathbf{X}\in\Omega} \mathbf{X} P(\mathbf{X})$, where Ω is the set that contains all the different states that $\mathbf{X}$ can be in. For example, if we would have only two places in the bag ($s = 2$) and only two words ($w = 2$), and 2 individuals then Ω would be:

$$\left\{ \begin{bmatrix} 2 & 2 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}, \begin{bmatrix} 1 & 2 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 1 & 2 \end{bmatrix}, \begin{bmatrix} 0 & 2 \\ 2 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 2 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 2 & 2 \end{bmatrix} \right\}.$$

We will collect the ensemble means of the word frequencies of individuals (i.e. the values of $\frac{x_{im}}{s}$) in a $n \times w$ matrix $\mathbf{F}$, with elements $f_{im} = \frac{1}{s} \sum_{\mathbf{X}\in\Omega} x_{im} P(\mathbf{X})$.

The ensemble means change over time as:

$$\begin{aligned} \frac{\mathrm{d}\mathbf{F}}{\mathrm{d}t} &= \frac{1}{s} \sum_{\mathbf{X}\in\Omega} \mathbf{X} \frac{P(\mathbf{X})}{\mathrm{d}t} \\ &= \sum_{\mathbf{X}\in\Omega} \sum_{i=1}^{n}\sum_{j=1}^{n} r_{ij} \left(\sum_{k=1}^{w}\sum_{m=1}^{w} \alpha_m \frac{x_{ik}+1}{s} \frac{x_{jm}}{s} \mathbf{X} P(\mathbf{X} - \mathbf{E}_{i,m} + \mathbf{E}_{i,k}) - \sum_{k=1}^{w}\sum_{m=1}^{w} \alpha_m \frac{x_{ik}}{s} \frac{x_{jm}}{s} \mathbf{X} P(\mathbf{X}) \right) \end{aligned}$$

We will next substitute $\mathbf{Y} = \mathbf{X} - \mathbf{E}_{i,m} + \mathbf{E}_{i,k}$ in the first sum, and denote with Ω' the set of all states the $\mathbf{Y}$ can be in:

$$\begin{aligned}
\frac{\mathrm{d}\mathbf{F}}{\mathrm{d}t} &= \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} r_{ij} \sum_{m=1}^{w} \alpha_m \frac{y_{jm}}{s} \left(\sum_{\mathbf{Y}\in\Omega'} \sum_{k=1}^{w} \frac{y_{ik}}{s} (\mathbf{Y} + \mathbf{E}_{i,m} - \mathbf{E}_{i,k}) P(\mathbf{Y}) - \sum_{\mathbf{X}\in\Omega} \sum_{k=1}^{w} \frac{y_{ik}}{s} \mathbf{X} P(\mathbf{X}) \right) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} r_{ij} \sum_{m=1}^{w} \alpha_m \frac{y_{jm}}{s} \sum_{\mathbf{Y}\in\Omega'} \sum_{k=1}^{w} \frac{y_{ik}}{s} (\mathbf{E}_{i,m} - \mathbf{E}_{i,k}) P(\mathbf{Y}) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} r_{ij} \sum_{m=1}^{w} \alpha_m \frac{y_{jm}}{s} \sum_{\mathbf{Y}\in\Omega'} \sum_{k=1}^{w} \frac{y_{ik}}{s} \mathbf{E}_{i,m} P(\mathbf{Y}) - \sum_{i=1}^{n}\sum_{j=1}^{n} r_{ij} \sum_{m=1}^{w} \alpha_m \frac{y_{jm}}{s} \sum_{\mathbf{Y}\in\Omega'} \sum_{k=1}^{w} \frac{y_{ik}}{s} \mathbf{E}_{i,k} P(\mathbf{Y}) \\
&= \sum_{i=1}^{n} \sum_{\mathbf{Y}\in\Omega'} \sum_{j=1}^{n} \sum_{m=1}^{w} r_{ij} \alpha_m \frac{y_{jm}}{s} \mathbf{E}_{i,m} P(\mathbf{Y}) \sum_{k=1}^{w} \frac{y_{ik}}{s} - \sum_{\mathbf{Y}\in\Omega'} \sum_{i=1}^{n} \sum_{k=1}^{w} \frac{y_{ik}}{s} \mathbf{E}_{i,k} P(\mathbf{Y}) \sum_{j=1}^{n} r_{ij} \sum_{m=1}^{w} \alpha_m \frac{y_{jm}}{s} \\
&= \sum_{\mathbf{Y}\in\Omega'} \frac{1}{s} (\mathbf{R} \cdot \mathbf{Y} \cdot \mathbf{diag}(\boldsymbol{\alpha})) P(\mathbf{Y}) - \sum_{\mathbf{Y}\in\Omega'} \frac{1}{s} (\mathbf{diag}(\mathbf{R} \cdot \mathbf{Y} \cdot \boldsymbol{\alpha}) \cdot \mathbf{Y}) P(\mathbf{Y}) \\
&= \mathbf{R} \cdot \mathbf{F} \cdot \mathbf{diag}(\boldsymbol{\alpha}) - \sum_{\mathbf{Y}\in\Omega'} \frac{1}{s} (\mathbf{diag}(\mathbf{R} \cdot \mathbf{Y} \cdot \boldsymbol{\alpha}) \cdot \mathbf{Y}) P(\mathbf{Y})
\end{aligned}$$

where $\boldsymbol{\alpha}$ is a vector with elements α_i.

The second term contains second order elements, and we can see that if there is selection of certain words the ensemble means do not form a closed system. If, however, all words are internalised with the same rate, as the results in the main paper suggest, and there is no selection between words we have $\alpha_m = \alpha$ for all m, then the ensemble means obey:

$$\frac{\mathrm{d}\mathbf{F}}{\mathrm{d}t} = \alpha \left(\mathbf{R} \cdot \mathbf{F} - \mathbf{diag}(\hat{\mathbf{r}}) \cdot \mathbf{F} \right),$$

where $\hat{\mathbf{r}}$ is a vector with elements $\hat{r}_i = \sum_{j=1}^{n} r_{ij}$. The ensemble means now are a closed system and change over time as a linear system of ordinary differential equations. As the results in the main paper suggest that internalisation rate is independent of word frequency, we will from hereon assume that all these rates are the same.

The ensemble mean of an individual's word frequency changes as:

$$\frac{\mathrm{d}f_{im}}{\mathrm{d}t} = \alpha \hat{r}_i \left(\sum_{j=1}^{n} \frac{r_{ij}}{\hat{r}_i} f_{jm} - f_{im} \right)$$

The word frequencies of a user change in response to other users in the network connected to this user, but are independent of the frequencies of other words. It is straightforward to solve this system of ODEs. However, to gain insight we will, rather than providing a general solution, solve a number of cases of special interest.

2. An individual exposed to constant word frequencies

As a first example we will study how an individual adjust its word frequencies to that of its environment. We will therefore assume that an individual communicates with a large number of individuals, and that the word frequencies in the

communication received are constant. Let the word frequency of word m that the focal individual i is exposed to be $h_{im} = \sum_{j=1}^{n} \frac{r_{ij}}{\hat{r}_i} f_{jm}$, which is assumed constant. We thus study the following w differential equations for the ensemble means of the word frequencies of the focal individual i:

$$\frac{\mathrm{d}f_{im}}{\mathrm{d}t} = \alpha \hat{r}_i (h_{im} - f_{im})$$

These systems have as equilibrium $f_{im} = h_{im}$, meaning that all word frequencies converge to the word frequencies that they are exposed to. The solution of the differential equations is

$$f_{im}(t) = h_{im} - (h_{im} - f_{im}(0))\, e^{-\alpha \hat{r}_i t}.$$

The word frequency of the words an individual is exposed to depends on the way that the individual is embedded in the social network. What is the effect of community structure on the word frequency if word frequencies transmit between communicators? Assume that individuals communicate predominantly with a subset of individuals which we call a community. By organising the network such that individuals' communication maximises the communication within communities, we can describe the network as a collection of communities. Such communities, it has been shown, tend to have distinct word frequencies (Bryden et al. 2013). Let the set c_k contain all the members of the k^{th} community, and that there are g such communities. The rate with individual i receives messages from members of c_k is $\hat{r}_i^k = \sum_{j \in c_k} r_{ij}$. In the messages that i receives from community c_k the word frequency of word m in the communication received is $h_{im}^k = \sum_{h \in c_k} \frac{r_{hi}}{\hat{r}_i^k} f_{im}$. The total word frequency received can now be partitioned as follows:

$$h_{im} = \sum_{k=1}^{g} \frac{\hat{r}_i^k}{\hat{r}_i} h_{im}^k.$$

The word frequency that a speaker will converge to in constant world is the weighted word frequency of all the communities that (s)he converses with.

If there is a dominant group in an individual's environment, and typically this would be the group that the focal individual is a member of, the word frequency of the dominant community will have most influence on the individual. This is compatible with the observations in Bryden et al. (2013), where it was found that 91% of Twitter conversation was within a community, and that the word frequencies of atypical words is shared within communities. It also is commensurate with the results of Tamburrini et al. (2015) where it is shown that outgoing messages from a community are more similar to the receivers, in terms of word frequencies than internal ones: if some members within a community communicate more with the outside world than others, those individuals will adjust their language more to the outside world then others. These individuals will be overrepresented in the outgoing communication. Outgoing messages from a group should therefore have word frequencies that are more like those of external groups than messages exchanged within the community.

There is one further aspect that this model can show. The word frequencies of an individual will change if the individual changes its environment, for instance, if it changes community. If that is the case there will be a change in the word frequencies received. Say that the new frequency of the word m that i receives is h_{im}. Let the word frequency of word m at the beginning of this process be equal to

$f_{im}(0)$ which is likely to be different from h_{im} because the individual was exposed to a different environment before. We will now describe how far the word frequency is removed from its final value.

The difference between the current word frequency, and the one that will be finally attained is $\Delta f_{im}(t) = h_{im} - f_{im}(t) = \Delta f_{im}(0)e^{-\alpha \hat{r}_i t}$. This leads to the observation that the relative convergence towards the frequency of the received messages is given by

$$\frac{\Delta f_{im}(t)}{\Delta f_{im}(0)} = e^{-\alpha \hat{r}_i t}.$$

What is remarkable is that the right hand side for this equation is independent of m: the relative convergence is the same for all words.

3. Two individuals converge over time, the more they communicate with one another

What if two individuals communicate? They will both change, yet also be subject to the environment that they are embedded in. If the word frequency coming from the environment is assumed to be constant, we have a system of two equations for the word frequencies of word m. At this point we will make our life easy and assume that $\hat{r}_1 = \hat{r}_2 = \hat{r}$, and $r_{21} = r_{21} = r$, for no other reason than that it simplifies the maths and makes the derivation easier to follow. It is not hard to relax this assumption and derive equivalent results. The frequencies of word m change over time as:

$$\begin{aligned}
\frac{\mathrm{d}f_{1m}}{\mathrm{d}t} &= \alpha\left((\hat{r}-r)h'_{1m} - \hat{r}f_{1m} + rf_{2m}\right) \\
\frac{\mathrm{d}f_{2m}}{\mathrm{d}t} &= \alpha\left((\hat{r}-r)h'_{2m} - \hat{r}f_{2m} + rf_{1m}\right),
\end{aligned}$$

where $h'_{im} = \sum_{h=3}^{n} \frac{r_{ih}}{\hat{r}_i - r_{i1} - r_{i2}} f_{hm}$, that is the word frequency of word m received by i from all individuals but individuals 1 and 2. It is further helpful to note that $\sum_{m=1}^{w} h'_{im} = 1$. This follows from the fact that $\sum_{m=1}^{w} f_{im} = 1$, which, in turn, follows from the fact that $\sum_{m=1}^{w} x'_{im} = s$.

To analyse this, we define $w_{1m} = (f_{1m} - f_{2m})/2$ and $w_{2m} = (f_{1m} + f_{2m})/2$. The equation for w_{2m} expresses how the overall frequency in the group community consisting if individuals 1 and 2 behaves. The w_{1m} equation shows how the two individuals converge towards each other. In these new variables the system changes as

$$\begin{aligned}
\frac{\mathrm{d}w_{1m}}{\mathrm{d}t} &= \alpha\left((\hat{r}-r)\frac{h'_{1m} - h'_{2m}}{2} - (\hat{r}+r)w_{1m}\right) \\
\frac{\mathrm{d}w_{2m}}{\mathrm{d}t} &= \alpha\left((\hat{r}-r)\frac{h'_{1m} + h'_{2m}}{2} - (\hat{r}-r)w_{2m}\right)
\end{aligned}$$

Now w_{2m} has an equilibrium at $(h'_{1m} + h'_{2m})/2$, which is simply the average of the frequencies received, and w_{1m} has an equilibrium at $\frac{(\hat{r}-r)}{\hat{r}+r}(h'_{1m} - h'_{2m})/2$. The solutions to these differential equations are:

$$\begin{aligned}
\frac{\hat{r}-r}{\hat{r}+r}\frac{h'_{1m} - h'_{2m}}{2} - w_{1m}(t) &= \left(\frac{\hat{r}-r}{\hat{r}+r}\frac{h'_{1m} - h'_{2m}}{2} - w_{1m}(0)\right)e^{-\alpha(\hat{r}+r)t} \\
\frac{h'_{1m} + h'_{2m}}{2} - w_{2m}(t) &= \left(\frac{h'_{1m} + h'_{2m}}{2} - w_{2m}(0)\right)e^{-\alpha(\hat{r}-r)t}
\end{aligned}$$

We can now reconstruct the solutions by using $f_{1m} = w_{2m} + w_{1m}$ en $f_{2m} = w_{2m} - w_{1m}$. The equilibria are for f_{1m} : $(\frac{\hat{r}}{\hat{r}+r}h'_{1m} + \frac{r}{\hat{r}+r}h'_{2m})$ and for f_{2m} : $(\frac{r}{\hat{r}+r}h'_{1m} + \frac{\hat{r}}{\hat{r}+r}h'_{2m})$.

The absolute convergence of the speakers is given by:

$$w_{1m}(t) - w_{1m}(0) = \left(\frac{(\hat{r}-r)}{\hat{r}+r}\frac{h'_{1m}-h'_{2m}}{2} - w_{1m}(0)\right)\left(1-e^{-\alpha(\hat{r}+r)t}\right)$$

The right hand side depends r: the bigger the r the faster you converge, but also the further you have to travel. So the result combines two components: the fact that the convergence is faster, and the fact that the end points are much closer together. Note that the right hand side also depends on m: this measure is different for different words.

One way of quantify how much two users differ in their language is through the use of appropriate measures. A widely used measure to assess similarity is the Bray-Curtis similarity measure. The Bray-Curtis similarity, for two identical sized bag of words is the sum over the lowest frequency frequency found in the bags. We can calculate this as

$$\sum_{m=1}^{w}\min(f_{1m},f_{2m}) = \sum_{m=1}^{n}\frac{f_{1m}+f_{2m}-|f_{1m}-f_{2m}|}{2} = \sum_{m=1}^{n} w_{2m} - |w_{1m}| = 1 - \sum_{m=1}^{n}|w_{1m}|.$$

The last step used $\sum_{m=1}^{n} w_{2m} = \sum_{m=1}^{n}\frac{f_{1m}+f_{2m}}{2} = 1$. The increase in the similarity measure over a period t is given by $\sum_{m=1}^{w} -(|w_{1m}(t)| - |w_{1m}(0)|)$, which is

$$\sum_{m=1}^{w}|w_{1m}(0)| - |w_{1m}(t)| =$$
$$\sum_{m=1}^{w}|w_{1m}(0)| - \sum_{m=1}^{w}\left|w_{1m}(0)e^{-\alpha(\hat{r}+r)t} + \frac{(\hat{r}-r)}{\hat{r}+r}\frac{h'_{1m}-h'_{2m}}{2}\left(1-e^{-\alpha(\hat{r}+r)t}\right)\right|.$$

To proceed we next assume that the sum of all words received, $\hat{r}$, exceeds the words received from the conversation partner. If $\hat{r} \gg r$ then the above expression simplifies to

$$\sum_{m=1}^{w}|w_{1m}(0)| - \sum_{m=1}^{w}\left|w_{1m}(0)e^{-\alpha(\hat{r}+r)t} + \frac{h'_{1m}-h'_{2m}}{2}\left(1-e^{-\alpha(\hat{r}+r)t}\right)\right|.$$

If we now define c_2 as the derivative of the above expression with respect to $e^{-\alpha rt}$:

$$c_2 = -e^{-\alpha\hat{r}t}\sum_{m=1}^{w}\left(w_{1m}(0) - \frac{h'_{1m}-h'_{2m}}{2}\right)\text{sign}\left(w_{1m}(0)e^{-\alpha(\hat{r}+r)t} + \frac{h'_{1m}-h'_{2m}}{2}\left(1-e^{-\alpha(\hat{r}+r)t}\right)\right).$$

If $w_{1,j}(0)$ and $m'_{1j} - m'_{2j}$ have the same sign this is a constant that does not depend on r. If the difference within most pairs is not too far from the point to which they will be eventually converge, then we can assume c_2 to be approximately constant. By defining

$$c_1 = \sum_{m=1}^{w}|w_{1m}(0)| - |\frac{h'_{1m}-h'_{2m}}{2}|$$

we now approximate the change in the Bray-Curtis similarity by

$$\sum_{m=1}^{w}|w_{1m}(0)| - |w_{1m}(t)| \approx c_1 + c_2 e^{-\alpha rt}. \tag{1}$$

4. Stochastic process for measuring incorporation rate

In our formalisation, we model each person as having an internal collection of words which are updated through communication with other people. The model is a Moran process where, with probability α, an encountered word (from another person) replaces a random word already in the internal collection. On Twitter we have data over a period of time of which words are received by a focal user and which words they broadcast. Consequently, we should be able to infer the incoporation rate α by modelling this internal collection as a hidden variable.

4.1. **Method.** The user's internal collection is defined as a collection of size s words. Many words are in the collection more than once with many repetitions of the more common words. For a word, we maintain a distribution of probabilities $p(i)$ $(0 < p < 1)$ that an individual has i representations of the word within their internal collection. Over time, we update the internal distribution according to words messaged to the target user from a second user and a proposed value of α. Following the logic of the Numerically Integrated State Space method (de Valpine and Hastings 2002), we use a procedure that generates a likelihood of the internal distribution given the outgoing messages of a user. Consequently, this method is able to calculate the extent to which the word usage of a second user has influenced the word usage of a target user.

The internal model is updated using incoming language through counting the number of mentions of the word (k) in the total number of N words in tweets directed at the user. The probability of n copies of the target word being internalised by the user is $Poiss(n, \alpha k)$, where $Poiss$ is the Poisson probability density function. Consequently, we can calculate the frequency of the first word internalised $\left(\sum_{n\geq 1} Poiss(n, \alpha k)\right)$, and, in general, of the mth word internalised $\left(\sum_{n\geq m} Poiss(n, \alpha k)\right)$. For each word internalised, we update the internal distribution as follows,

$$\begin{gathered}\forall i \in \mathbb{Z}, 0 \leq j < s \\ \forall m \in \mathbb{Z}, m < j < s \\ p(i) = p(i,j) + \sum_{n\geq m} Poiss(n, \alpha k) p(i-1)(1-(i-1)/s) \\ - \sum_{n\geq m} Poiss(n, \alpha k) p(i)(1-i/s)\ ,\end{gathered}$$

taking $p(-1) = 0$ to resolve the boundary. Following this process, for the $N - k$ words that are not the target word, we use a similar procedure to update the internal model.

$$\forall i \in \mathbb{Z}, 0 \leq j < s \tag{2}$$

$$\forall m \in \mathbb{Z}, m < j < s \tag{3}$$

$$p(i) = p(i) + \sum_{n\geq m} Poiss(n, \alpha k) p(i+1)(i+1)/s \tag{4}$$

$$- \sum_{n\geq m} Poiss(n, \alpha k) p(i) i/s\ . \tag{5}$$

Given the internal model of a user's propensity to use a word, we are able to generate the probability of the language they generate at a point in time t. For a user who has used k mentions of a specific word in N words, we calculate the probability of the internal model for the word as,

$$Pr_t(k, N) = \sum_i p_t(i) B(k; N, i/s),$$

where B is the binomial probability density function. The internal model can now be updated to reflect this new information for the next time point.

$$\forall i, p_{t+1}(i) = \frac{p_t(i) B(k; N, i/s)}{\sum_i p_t(i) B(k; N, i/s)} \,. \tag{6}$$

We can now calculate the likelihood of the model and parameters (H) given the data (D) as,

$$\mathcal{L}(\mathcal{H}|\mathcal{D}) = \prod_t Pr_t(k, N) \,. \tag{7}$$

4.2. **Tests.** The method was tested by taking two users with different internal frequencies of a specific word. At each time segment, both users output a random number of words with the frequency of the specific word picked from a binomial distribution according to their internal frequency of the word. One of the users is the target user, who receives languague from the other, and updates their internal model according to a hidden value of α. We then run the method against the language generated by the two users to generate find the value of α with a maximum likelihood (Eq. 7) and compare that against the hidden value - see Fig. 1.

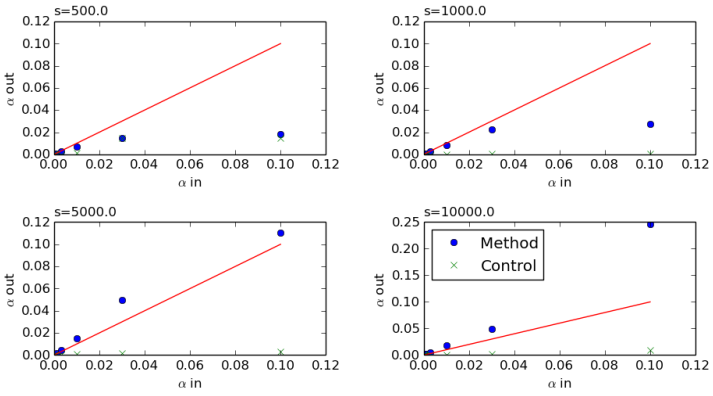


FIGURE 1. Four different values of s were tested. At low s, and higher levels of α, the target user quickly adopted the other's usage. At the highest level of $s = 10,000$, the value of α was over estimated.

4.3. **Implementation.** We looked at pairs of users, a *target* user for whom we maintain an internal collection, and an *incoming* user from whom we monitor their incoming tweets. To some extent, two users will mirror one another in conversation, and so we would expect to see some transitory usage of the same words. We can

account for this by ignoring tweets where a target user sent messages directed back to the incoming user. We then looked at time segments where both the outgoing and the incoming user posted.

The initial condition of the internal collection was set at a binomial distribution, the mean of which is calculated by the relative frequency for the specific word over the first half of the time segments (multiplyed by s). There was then a burn in period for half the time segments where α was set to 0.0. At this point the internal frequency should reflect that of the usage of the target user. We then introduce update of the internal model from incoming language (Eq. 5). A control was generated where the outgoing tweets of the target user were randomly shuffled so that the time signal was lost.

We focussed on 1,000 words randomly sampled from all the word instances (so including copies of words) we had in our complete sample of Twitter. We looked at 10,000 pairs of users, where both the users had tweeted at least one tweet to each other and at least 500 conversational tweets to all users. For a given value of α we calculated the log-likelihood by summing the log-likelihoods for each pair of users. For each word, we then used Nelder-Mead optimisation (Nelder and Mead 1965) to find the value of α with the maximum likelihood.

We found, when running optimations, that increasing the level of s in the process increased the optimal level for α. When we update the internal model according to output from the target user (see Eq. 6), the amount of change is independent of the value of s. However, when we update the internal model due to incoming language (Eq. 5), the effect of this is dependent on the value of s. To correct for the increase in α needed to compensate for this effect, we subtracted the value of α with the maximum likelihood for the control from the value of α generated with the unshuffled time signal.

5. References

- Blythe, R. A. (2012). Neutral evolution: a null model for language dynamics. Advances in Complex Systems, 15(03n04), 1150015.
- Blythe, R. A., & McKane, A. J. (2007). Stochastic models of evolution in genetics, ecology and linguistics. Journal of Statistical Mechanics: Theory and Experiment, 2007(07), P07018.
- Bryden, J., Funk, S., & Jansen, V.A.A. (2013). Word usage mirrors community structure in the online social network Twitter. EPJ Data Science, 2:3, 1-9.doi:10.1140/epjds15.
- de Valpine P. & Hastings A. (2002). Fitting population models incorporating process noise and observation error. Ecological Monographs, 72:5776.
- Nelder J.A & Mead R. (1965). Simplex Method for Function Minimization. The Computer Journal, 7:30813.
- Tamburrini, N., Cinnirella, M., Jansen, V. A., & Bryden, J. (2015). Twitter users change word usage according to conversation-partner social identity. Social Networks, 40:84-89.

5. Resurgent Insurgents: Quantitative Research Into Jihadists Who Get Suspended but Return on Twitter

Overview in relation to the thesis

The previous chapter has shown that interaction and communication can lead to horizontally inherited changes in people's language; an important assumption for analysing language used on Twitter profiles as proxy for underlying patterns. Having validated the approach, this chapter can begin to investigate social phenomena; in particular, a phenomenon that arises when analysing supporters of terrorism on Twitter.

Scholars of terrorism have investigated and characterised users expressing extremist views or support for terrorism as such users are very active on Twitter. Their accounts, however, frequently get suspended for violating Twitter's terms and conditions. A debate over the effectiveness of this suspension has, therefore, arisen (Arthur, 2014; Fisher, 2015; Gladstone, 2015; Stern and Berger, 2015). An important factor in the debate is the ease and speed with which new accounts can be created. Users thus resurge back onto Twitter like the moles in the “whack-a-mole” game (Arthur, 2014; Berger and Morgan, 2015; Levy, 2014; Stern and Berger, 2015). Multiple accounts belonging to the same user are problematic as they can cause biases, such as pseudoreplication, for analysts of terrorism and intelligence. Whilst the existence of resurgents is widely acknowledged, little research investigates the problem; conclusions

about them have been drawn from case studies on one (Stern and Berger, 2015) or four (Berger and Perez, 2016) users.

This chapter uses novel methods, based on observations about the similarity between multiple incarnations of terrorist accounts (Barnett, 2013; Stern and Berger, 2015), to identify resurging accounts from a sample closely linked to jihadist terrorists and provides detailed analysis going beyond previous case-studies. It shows that suspension is less disruptive to terrorists than previously thought, but that the bias and disruption caused to terrorism research has been underestimated (Wright *et al.*, 2016).

Wright, S., Denney, D., Pinkerton, A., Jansen, V.A.A., Bryden, J., 2016. Resurgent Insurgents: Quantitative Research Into Jihadists Who Get Suspended but Return on Twitter. *Journal of Terrorism Research*. **7**(2): 1–13. DOI: http://dx.doi.org/10.15664/jtr.1213

The Centre for the Study of Terrorism and Political Violence
Journal of Terrorism Research

Articles

Resurgent Insurgents: Quantitative Research Into Jihadists Who Get Suspended but Return on Twitter

by Shaun Wright, David Denney, Alasdair Pinkerton, Vincent A.A. Jansen, John Bryden

Abstract

Jihadists are very active on Twitter but their accounts frequently get suspended. A debate over the effectiveness of suspension has arisen; an important factor is that Jihadists quickly create new accounts, resurging back like the moles in the "whack-a-mole" game. This causes biases for terrorism and intelligence analysts. Whilst widely acknowledged, little research investigates the problem. In this study we identify resurging Jihadist accounts with novel methods, and provide detailed analysis going beyond previous case-studies. We show that suspension is less disruptive to terrorists than previously thought, whilst the bias and disruption caused to terrorism research has been underestimated.

Introduction

Jihadists have taken to social media. Twitter has emerged as their "favourite" site (Weimann, 2014) and an estimated 46,000-90,000 ISIS supporting accounts were active there in Autumn 2014 (Berger and Morgan, 2015). Jihadists use Twitter for a variety of reasons. The first reason is to spread their messages to a wide audience. The second is for recruitment; the third is to indoctrinate further those drawn to them, like a crucible of radicalisation. And finally, (although not comprehensively) they also use Twitter for seemingly mundane conversation amongst friends.

As a consequence of the volume of data, and its open-source nature, analysis of this source of intelligence about terrorist and extremist activity is becoming more common amongst academics, journalists and government practitioners (Chatfield, 2015; Greene, 2015; Magdy, 2015; Mahmood, 2012; Moriarty, 2015; Ryan, 2014; Stern and Berger, 2015). Whilst there is very detailed research on the Twitter structure and strategies of the top-down, officially-controlled tiers of Jihadist terrorist groups (Stern and Berger, 2015), we argue that the field could benefit from more sustained research on the larger, bottom-up community of Jihadist massed ranks.

Another consequence of how numerous and vocal Jihadists are on Twitter, is the political, cultural and media pressure to take down – or suspend – terrorism supporting accounts (Levy, 2014; Moriarty, 2015). In recent years this has led to several changes in Twitter's suspension policy, and an enormous increase in the number of suspensions. A debate has now arisen in the media and academic literature on the effectiveness of these suspensions (Arthur, 2014; Fisher, 2015; Gladstone, 2015; Stern and Berger, 2015). The assumption is that suspending terrorist supporting accounts reduces the number of terrorists on Twitter. It is assumed that this, in turn, will help counter the objectives for which Jihadists are using social media in the first place:

recruitment, radicalisation, spreading propaganda and threats. On the other side of the debate are concerns over loss of intelligence, freedom of speech, and how realistically achievable the number of suspensions needed to make a dent in the problem is.

Central to this debate is another significant and problematic phenomenon associated with Jihadist social media research: "many of those suspended users simply sat down at their computers the very next day, created new accounts, and started all over again" (Stern and Berger, 2015). This phenomenon is acknowledged in a range of studies (Chatfield, 2015; Magdy, 2015; Berger and Morgan, 2015) and widely referred to as "whack-a-mole" (Arthur, 2014; Berger and Morgan, 2015; Levy, 2014; Stern and Berger, 2015). Those who create these resurging whack-a-mole accounts we call "resurgents" and we provide a more detailed definition later in the paper.

Resurgents do not just cause a whack-a-mole challenge for those performing the suspensions. Their quantity makes identification difficult and so they often go unnoticed. The impact of researchers being unable to identify or control for resurgents is that their datasets will suffer biases; the main bias being replicate error. If the dataset contains duplicate resurgent accounts who get treated as independent data points, this clearly causes errors in any research addressing a range of issues: the number of Jihadist accounts, the level of support for a particular course of action, how unusual a particular behaviour is, and so on.

An example of a problem caused by resurgents is Berger and Morgan's estimate of the (carefully worded) number of "ISIS-supporting Twitter accounts". The problem is that we do not know how many unique ISIS supporters are represented by these accounts. In another example, Twitter claimed that it had suspended 10,000 ISIS linked accounts in a single day (Gladstone, 2015). Again, it is unknown how many ISIS supporters this represents. These problems occur because there are no methods to identify resurgent accounts amongst this volume of data, or control for the biases that they cause. One of our aims is to help develop such methods and provide these estimates.

It is clear that resurgents cause problems for suspension and for research, yet academic study of them is lacking. Previous studies have discussed suspension and resurgence as a potential flaw with the generalisability of their findings (Berger and Morgan, 2015; Chatfield, 2015; Magdy, 2015). However, almost no research has been done to characterise and describe suspended or resurgent accounts – partly due to the lack of methods for finding them. The impact of resurgents on the effectiveness debate, therefore, currently rests on Stern and Berger's (2015) case study of a single resurgent.

Stern and Berger (2015) conducted a case study of the suspension and single resurgence of the official al Shabaab Twitter account in January 2013 and concluded that suspension is disruptive to terrorists but not to research or intelligence gathering. One of their claims is that finding matching resurgent accounts, and analysing them as continuations of the same account is easy. Furthermore, they claimed the "suspension had cost nothing in intelligence value... and the new account continued the stream of press releases". Whilst this may be true for researchers tracking a particular case study account, especially official media accounts, any researcher analysing the Jihadist massed ranks on Twitter is going to struggle. We suggest that trying to identify all corresponding resurgent accounts in a dataset of 46,000-90,000 accounts is so time-consuming for humans that there *is* likely to be an intelligence cost. Addressing this hypothesis is another one of our aims in this paper.

Stern and Berger also determined the rate at which their resurgent case-study account accrued followers and calculated that it would take months or years to regain all their followers. They then argued that suspension imposes "clear numeric costs" since ISIS supporters must "reconstruct their social networks and reestablish

trust" (Stern and Berger, 2015). While there may be costs for some suspended accounts, this picture is incomplete. We hypothesise that because Jihadist accounts have previously (and repeatedly) built their reputation and the trust of the community, when they return as resurgents, the nature of Twitter means that they can quickly seek out close comrades from their previous network, initiate contact and re-establish their credentials. Therefore we predict that the number of followers of resurgent accounts should grow faster than naturally growing Jihadist accounts who must establish credentials from scratch rather than simply renew them.

We will also consider other factors that could explain any accelerated growth amongst resurgents. One relevant Twitter phenomenon could be "Follow Friday" (Leavitt, 2014), where participating users recommend accounts (on Fridays) to their followers. These tweets are often signposted with the hashtags "#ff" or "#followfriday", e.g. *"#ff #followfriday @randomuser1 @twitteruser123"*. We hypothesise that they could be helping to drive growth, and will perform an initial test of how common they are to assess the viability of this.

We think that the phenomenon of accounts resurging from suspension is a significant enough feature of modern terrorism to merit further study and definition. With currently only a single case study, we suggest that the next logical step is to study more resurgents, and this is the main aim of our paper. However, since the world of modern terrorist activity is one of social media and big data, conclusions drawn about case studies cannot be appropriately generalised to the whole population of Jihadists. We therefore, as has been identified as necessary in the study of Twitter Jihadists in general, propose using big data methods (Berger and Morgan, 2015) on a large sample of resurgents.

We define a Twitter resurgent as any user who has created multiple accounts on Twitter under different handles (unique user-names beginning '@'). Resurgence does not only occur as the direct result of suspension; some users pre-empt their suspension by changing their handle or operating multiple backup accounts. All resurgent types are included in the definition, however, as they cause the same biases to research datasets. On the other hand we do exclude those who are consciously masquerading as different people (e.g. operating multiple personas or a variety of automatic bots) and we consider the implications of this in the discussion.

In this paper we aim to find sets of accounts belonging to the same resurgents. Once we have done that, we can study and describe them. We will assess how disrupted they are by quantitatively analysing the rate at which they accrue followers compared to non-resurgent accounts, as well as looking at Follow-Friday as a possible driving mechanism. We will also provide an estimate of the proportion of Jihadist accounts which are just duplicates and the proportion which represent unique Jihadists. These findings will give terrorism researchers a better understanding of the true numbers and distribution of Jihadists on social media, as well as an appreciation of how disruptive suspension is for research. We therefore set out the first large scale description of resurgent Jihadists, a significant phenomenon in modern terrorism, challenging, in the process, some of the conclusions about Jihadist social media behaviour drawn by others.

Methods

1. Dataset

The sampling algorithm used was developed to bias sampling toward accounts that tended to have numerous links to other accounts that had already been sampled. The reason for doing this was the principle

of homophily: the tendency of people to associate with others similar to them (McPherson, 2001). This principle has been shown to lead to highly intra-linked communities on Twitter that bias their interactions to other members of the community and share a social identity (Bryden, 2011; 2013; Tamburrini, 2015). Consequently, we reasoned that Jihadists would bias the accounts that they followed towards other Jihadist accounts and set up our sampling algorithm accordingly.

We therefore used weighted snowball sampling (Goodman, 1961) to identify Jihadist Twitter accounts. This approach enabled us to grow the sample, whilst weighting sampling towards accounts with numerous links to accounts already identified. A handful of publicly-known, official "media" Jihadist Twitter accounts named by newspapers provided our starting point. We then manually inspected the Twitter followers of these accounts, aided by Twitter's "Who to follow" algorithm, and from our analysis we identified 34 'unofficial-but-supporting', Jihadist accounts. For practicality, we selected only English speaking accounts. We then used this starting sample to seed the snowball algorithm.

We snowball sampled daily between May and July 2015 (77 days, with power issues preventing sampling on 10 days). On each day we looked at all accounts followed by those already in our sample. We then sampled any account identified as being followed by >10% of the users in our sample, and with <1,000 followers of its own.

We selected the 10% threshold to grow the sample slowly, without accelerating, whilst remaining within a relatively tight community of English speaking Jihadists (the principle of homophily). While our sample was smaller than 100 users we used a fixed threshold (new accounts must be followed by more than 10 accounts in our sample). We switched to the 10% threshold once we had sampled 100 accounts.

The upper limit of 1,000 followers was selected for two reasons. Firstly, to prevent the inclusion of popular journalists and academics who are often both highly interlinked with the networks, and connected outwards to non-Jihadist followers. Such community transcending journalists were liable to divert the sampling away from the Jihadist community. Our cut-off is similar to, although more ruthless than, the precedent set by Berger and Morgan (2015) who used a 50,000 cut-off, finding that accounts more popular than this were unrelated. Secondly, by avoiding the more 'popular' accounts, we aimed to direct our dataset away from the official, top-down Jihadist media accounts covered in other research, and towards the largely neglected Jihadist massed ranks.

During sampling, some accounts were protected, suspended or had voluntarily changed their user-name. We moved these to an "inactive sample" where we recorded all the account information, but they no longer contributed to the 10% threshold check. We identified suspended users by the official suspension report with which Twitter had replaced their pages. Protected users had activated privacy settings and only biography, pictures and summary meta-data were available. Non-existent accounts display an official Twitter message that the user cannot be found (despite our evidence that they previously did). Although no information is provided about their non-existence, since Twitter does not report them as suspended we assume that the users changed their handles themselves.

2. Finding resurgent Jihadists

To identify resurgent accounts we used a quantitative approach that helped draw our attention to accounts whose Twitter biographies, names and locations contained at least 30% of the same words. We set out the rationale for why our quantitative approach is needed, over human identification of accounts, in

Supplementary Material 1. We then visually assessed those accounts, identifying and classifying resurgents.

Defining a set of accounts belonging to a resurgent

When comparing accounts, we used open criteria for determining whether they formed a set. However, in practically all cases, an almost identical match between handle, name, biography or location was necessary and sufficient. Biography and handle were the strongest indicators, whilst location, surprisingly, was still informative due to peoples' unique spelling, punctuation, and choice of descriptive terms. A hypothetical, illustrative example of an almost identical match would be the handles "@jihad_bob2" and "@jihad_bob3".

Figure 1. An illustrative example from our data of two resurgent accounts which we classified as a set. They have almost identical handles and almost identical biographies. Their images were not inspected, but their profile images are an almost identical match too. Screenshots of two user accounts taken from http://twitter.com.

Figure 2. An illustrative example from our data of two accounts which we did not classify as duplicates of one another, despite some similarities. Screenshots of two user accounts taken from http://twitter.com.

3. Do resurgents accrue followers faster?

We investigated how disrupted resurgent accounts are by calculating their rate of follower accrual versus non-resurgent controls. As we were unable to find other matching resurgents, we treated all those who had not been identified as non-resurgent controls. We calculated growth rate by dividing the number of followers an account had upon sampling by the number of days between creation and sampling. We used the non-parametric Mann-Whitney U test after ruling out normality (both p = 0.00, 2.d.p, Kolmogorov-Smirnov).

Results

We sampled 1,920 English speaking Jihadist accounts from Twitter. By the end of sampling 1,080 had been suspended, 141 accounts were private, 97 no longer existed and 602 were active (figure 3). Only 1,858 of the users had sufficient name, location and biography information for analysis.

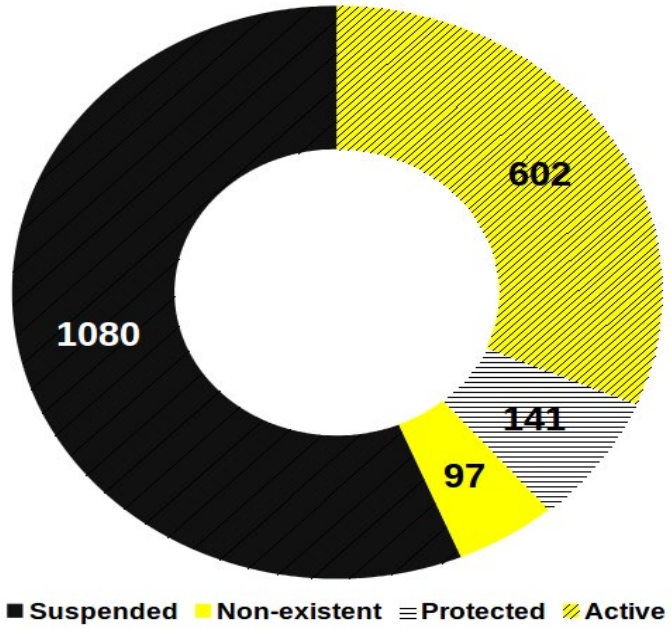


Figure 3. The distribution of our dataset of 1,920 English speaking Jihadist Twitter accounts. By the end of sampling, 1,080 had been suspended by Twitter, 141 had set their accounts to private, 97 no longer existed due to voluntary name change and 602 were still active.

1. Terrorist group affiliations

The majority of accounts do not declare a terrorist organisation affiliation, nor does a simple content analysis allow for unequivocal categorisation. Amongst 300 randomly selected users, 39 (13%) provided an allegiance, of which all gave ISIS, IS, Islamic Caliphate, Baqiya or Khilifa. Amongst the 261 that didn't, 34 (13%) gave one of the four most common locations: "Dar ul Kufr" [Land of the unbelievers] (n=16, 6%), "UK" (n=12, 5%), "Dunya" [the non-spiritual, temporal world] (n=3, 1%) , and "Somalia" (n=3, 1%); with the sharing of extremist content and pro-Caliphate sentiment also common. Twitter also suspended 56.3% of our sample, evidence that suggests they were engaging in extremist activity. We therefore categorise our sample as Jihadists, whilst assuming, based on location and content, that the majority are ISIS-supporting members of the "Baqiya family" (Amarasingam, 2015).

2. Finding resurgent Jihadists

Using the quantitative approach outlined in the methods, we estimated the number of unique Jihadist users by identifying resurgents: users in the dataset who had multiple, matching replicate accounts.

From 1,858 user accounts with information to analyse, only 1,484 (79.9%) were unique Jihadists. The remainder, over one in five accounts, were duplicates: resurgent accounts. 192 (12.9%) of the unique users were resurgents who owned, on average, 2.95 accounts (a set of mean size 2.95) within the three month period (table 1).

The other statistic commonly reported is the number of Jihadist accounts that have been taken down or suspended. This also overestimates the number of unique Jihadists. Performing the same analysis with the suspended users with information to analyse (n=1,066), we found only 757 (71.0%) unique Jihadists. 114 (10.7%) of these unique users were resurgents, owning a mean of 3.71 suspended resurgent accounts in three months (table 1).

	Number of accounts analysed	**Number of unique Jihadists**	**Number of duplicate accounts**	**Number of unique users who were resurgents**	**Mean number of accounts belonging to each resurgent**
Entire sample	1,858	1,484 *(79.9%)*	374 *(20.1%)*	192 *(12.9%)*	2.95
Suspended users	1,066	757 *(71.0%)*	309 *(29.0%)*	114 *(10.7%)*	3.71

Table 1. Identification and quantification of resurgents in the dataset: users who had multiple, matching replicate accounts.

3. Resurgents accrue followers faster

We found that the growth rate of resurgent accounts (n=566, median 43.8) is significantly greater ($p < 0.0001$ [exact p-value $< 2.38 \times 10^{-40}$], 1-tailed Mann-Whitney U) than that of naturally growing, non-resurgent accounts (n=1,292, median 8.37) (figure 4).

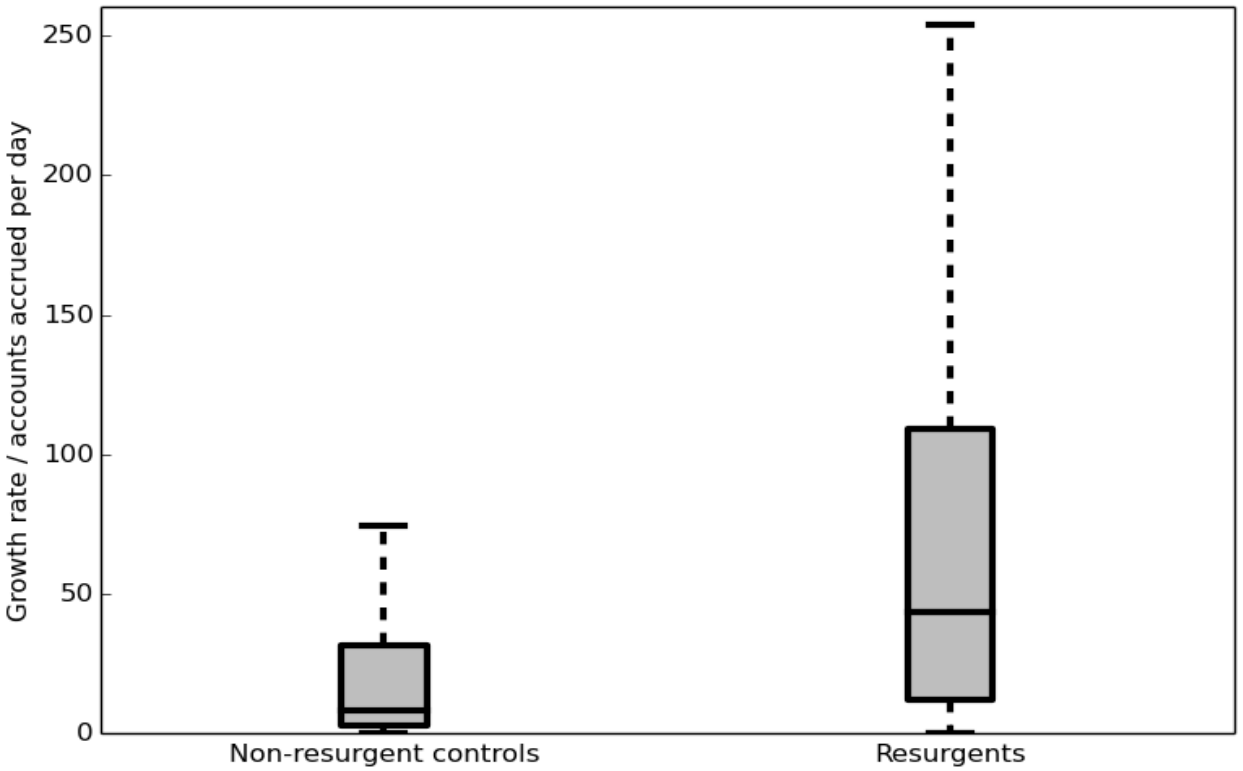


Figure 4. The growth rates (followers accrued per day) of resurgent accounts (n=566) versus naturally growing Jihadist accounts (n=1,292). The plot shows the growth rate of resurgent accounts is significantly higher than that of non-resurgent accounts.

4. Jihadist Follow-Friday

Having shown that resurgent accounts grow faster than those of non-resurgents, we searched for explanatory factors. We observed a similar phenomenon to "Follow-Friday" within the Jihadist Twitter community and assessed their viability as a growth driving mechanism by testing how common these tweets were.

Downloading the entire daily tweet output of our sample generated a corpus of approximately 155,000 tweets. We randomly-selected 2,500 tweets from this corpus; 46 (1.84%) fitted the Jihadist Follow-Friday structure.

Although we dub them "Jihadist Follow-Friday" tweets, zero (0.0%) contained Friday hashtags. Furthermore, none (0.0%) of the 46 examples promoted more than one user per tweet, with 17 (37.0%) repeating the name several times per tweet, e.g. *"Follow: @jihadistaccount123 @jihadistaccount123 @jihadistaccount123"*, and the remaining 29 (63.0%) naming them only once, e.g. "*FOLLOW & SUPPORT @jihadistaccount123*". Three tweets (6.52%) also stated that the user had returned from suspension.

As an indicator of whether Jihadist Follow-Friday tweets are significant enough to contribute to re-growth, this result estimates that there are 2,852 tweets (1.84%) promoting other Jihadist accounts in our dataset of 155,000 tweets.

Discussion

Suspension and resurgence are significant phenomena in modern, online terrorism. As resurgents are difficult to find in large numbers, research into them is scarce, relying on Stern and Berger's (2015) case study alone. Furthermore, terrorism research treats the duplicate resurgents as independent data points,

biasing social media research into the numbers, opinions and behaviour of Jihadists. We found resurgents, estimating that within our sample only 79.9% of Jihadist Twitter accounts belong to unique Jihadists, with a lower 71.0% of unique Jihadists amongst suspended accounts. This gives researchers a better picture of the patterns displayed by resurgents, as well as a scale of the significant biases for research and estimates and the continuous disruption to intelligence gathering.

With the identification of resurgents comes the ability to analyse them beyond individual case studies. Previous work has concluded that there are "clear numeric costs" to resurgents who suffer slow regrowth as a cost of suspension (Stern and Berger, 2015), contrary to this single al-Shabaab account however, we have shown that in our sample resurgents grow significantly faster (median 43.8 accounts accrued per day) than non-resurgent Jihadists (median 8.37). Whilst it remains possible that this might not be sustained long enough to get back all of their old followers, especially the curious Westerners, there is no obvious disruption to Twitter when considered as a crucible of radicalisation. Whether or not Jihadist Follow-Friday tweets help to drive this accelerated growth also merits further study, as they seem prominent (1.84% of tweets) given the number of alternative discussion topics.

Our findings could help analysts to put reported numbers and statistics in a more appropriate context. Berger and Morgan estimated the number of ISIS supporting Twitter accounts at 46,000-90,000. However, we have shown that an improved estimate should drop below 36,800-72,000 (79.9%) unique users. Another commonly reported, headline-catching statistic is the number of ISIS accounts suspended; Twitter reported suspending 10,000 accounts. However, our results suggest that this should be corrected to represent only 7,100 (71.0%) unique ISIS supporters. We suggest that while the rate of suspensions remains stable, our specific results of 79.9% (overall) and 71.0% (amongst suspended) may have some usefulness, but that even when suspensions escalate, the principle behind our finding remains crucial. All of these results highlight the dangers in working with a Jihadist dataset without correcting it for resurgents.

One of the implications of this improved picture of resurgents is the contribution to the suspension effectiveness debate. A great deal of political and public pressure exists to suspend terrorists and their supporters from social media sites. Although intelligence concerns often take "a distant third" place to business and cultural concerns, some argue that the intelligence costs are limited (Stern and Berger, 2015). Whilst our results do not address the cultural or ethical arguments, they do suggest that suspensions are less disruptive to terrorists than previously argued; furthermore, suspensions cause significant biases to data and its analysis. Rather than leading us, however, to advocate against suspension – there are convincing ethical and intelligence quality improving arguments (Stern and Berger, 2015) – we propose using methods to control for it.

We consider our dataset of accounts, and their suspension rates, to be generalisable to the unofficial, English-speaking, Jihadist community on Twitter. We categorised our sample as pro-ISIS members of the "Baqiya family" (the friendly network of online ISIS supporters) (Amarasingam, 2015), although terrorist group affiliation is almost impossible to assess without additional sources of data. It is, however, in line with the political dominance of ISIS during summer 2015, the nature of the "Baqiya family" (Amarasingam, 2015), and Berger and Morgan's (2015) estimate of 46,000-90,000 ISIS-supporting accounts during a similar length time. Although it is possible that generalisability is limited by snowball sampling's bias towards the seed list, after sampling 1,920 accounts from a seed list of 34, any initial bias should have been diluted. We therefore associate our results only with the general "Jihadist" community, limiting the ability of our study to make statements about differences between specific terrorist groups. Inspection of the data does, however, indicate

success in our aim of using a minimum popularity to exclude bots.

A potential critique of our sampling method (continually looking for new accounts) is that it could be biased towards resurgents. We defend the appropriateness of our sampling, however, as it will still snowball into a wider community, reaching out to newly discovered accounts that need not be new to Twitter. We would also point out that although snowball sampling cannot reach disjoint groups, such a hypothetical, unconnected terrorist group is by definition unrepresentative of the ISIS-dominated Twitter environment. We do, however, suggest that the best course of action is for researchers themselves to analyse their dataset for resurgents. Finally, our definition of resurgents also excludes those masquerading as bots or multiple personas. These are phenomenon potentially causing additional replicate biases to terrorism research and therefore merit further research.

Although there appear to be some statistical issues with generalising our findings directly to Berger and Morgan's work, there are several possible counter-explanations. Scaling by 79.9% predicts that over 20% of their users have resurged back, but they only reported ~7.5% being suspended in the first place. There are however, three reasons why this need not contradict our findings, nor stop us applying our result to their data. Firstly, they acknowledge that the suspension rate has dramatically escalated since, and in our data it was 56.3%. Secondly, name-changing and backup accounts are also sources of resurgents and are presumably not covered under their reported suspension statistics. Finally, it appears that their sample was not continuously re-checked for suspensions. Thus their suspension rate may actually be higher than reported. In the specific case of our Twitter example, where all the accounts were active during a single day, our findings may also not be applicable. However, whenever accounts are reported suspended over a period greater than several weeks, our findings may be highly informative. Again, these challenges only emphasise the importance of researchers attempting to find resurgents in their data for themselves.

Our study included several types of resurgents, including backup accounts and those created after suspension. The difference between a backup and post-suspension account is not a binary classification but a spectrum, depending on whether the main account has been suspended, the age of the backup before and since becoming the main account, and the ratio between these. Recording data to investigate these is therefore beyond the scope of this article, but merits a future study. Crucially however, a lower rate for backups would lower the rate for combined resurgents, and this thus indicates the robustness of our significantly elevated result.

Limitations

A limitation of our "Jihadist" study is that we cannot make statements about the differences between specific terrorist groups. These findings could also benefit from more work with a broader sampling procedure, as there are limits on generalising our sample to the unofficial, English-speaking, Jihadist, Twitter community (snowball sampling methods both limit the ability to reach disjoint groups, and exhibit bias towards their seed lists). Additionally, our estimates are conservative upper bounds as we could have missed some resurgents due to the challenge of finding resurgents amongst big data. Our estimates are also upper bounds as our definition excluded those masquerading as bots or multiple personas, and our study amalgamated several types of resurgents, including backup accounts and those created after suspension. Although there are likely to be differences between backup and post-suspension resurgent accounts (we hypothesise that their longer lifespan and insignificance to followers would give backup resurgents a lower growth rate), testing this is non-trivial. There may also be limitations with generalising our findings directly to all other numerical

estimates, as sampling methods differ from study to study.

Conclusion

This paper marks a step change in methodological approaches towards the study of resurgent Jihadists. The new methods give us novel insights into the proportion of fast-growing, duplicate accounts (20-30%), which in turn suggest some crucial new approaches in terrorism studies: adjusting numerical estimates, recognising dataset biases, and seeking methods to identify and control for the significant number of resurgents. Our quantitative method in particular, which we hope to calibrate further in future work, appeared very useful for quickly finding resurgents, and this presents a clear example of the wider importance and power of using quantitative analysis to investigate a range of terrorism behaviours.

Author contributions

All authors were involved in the conception of the work. SW collected the data, performed the analyses, wrote the first draft and led the writing of the manuscript. DD, AP, VAAJ and JB edited and critiqued the manuscript. The authors would also like to express their gratitude to Peter Adey for helpful discussions and feedback on the manuscript, and to the reviewers for their constructive comments. Date submitted: 19 October 2015; Accept Submission: 07 March 2016.

About the authors

Shaun Wright *is a PhD candidate at Royal Holloway University of London where he is also a visiting teaching assistant and guest lectures on the Terrorism, Insecurity and Risk course. Reflecting the interdisciplinary nature of terrorism studies, he has supervisors in the departments of Biological Sciences, Law, Criminology & Sociology, and Geography. He studied for his undergraduate BA at Gonville & Caius College, Cambridge University, combining courses in Computer Science, Psychology and Neuroscience.*

David Denney *is Professor of Social and Public Policy in the School of Law, Royal Holloway University of London. He has written extensively on various aspects of risk in society, violent crime and discrimination in the criminal justice system. He is currently Principal Investigator on ESRC Dstl funded research which examines the impact of social media on military personnel and their families.*

Alasdair Pinkerton *is a political geographer with particular interests in issues related to critical geopolitics, the media, and the 'international relations' of public diplomacy. He has regional specialisms in the South Atlantic and South Asia, as well as working extensively in the US, Canada, the UK and Cyprus. He appears frequently in the UK and international media on issues related to global geopolitics.*

Vincent Jansen *is a Professor of Mathematical Biology at Royal Holloway University of London. He is interested in population dynamics and evolution, including the dynamics and evolution of social systems. His research entails the formulation and analysis of mathematical models to understand biological processes, and he has applied this to a range topics in biology and beyond. Part of his research has addressed how social networks form, how people communicate on social networks and how our language is shaped and influenced by the social network that we are embedded in.*

John Bryden *is a Research Fellow at Royal Holloway University of London interested in group behaviour, dynamics and evolution in biological systems and human societies. More generally, he is interested in the*

modelling of complex systems and the methodological issues raised by the modelling of complex systems. He is building a methodological approach for the studying of human behaviour in social systems by applying models to large data sets. He has recently published work studying how humans cluster into groups on social networks and the effects this has on their language.

References

Amarasingam, A., 2015. What Twitter Really Means For Islamic State Supporters. *War on the Rocks*. December 30.

Arthur, C., 2014. Taking down Isis material from Twitter or YouTube not as clear cut as it seems. *The Guardian*. June 23.

Berger, J.M., Morgan, J., 2015. The ISIS Twitter Census: Defining and describing the population of ISIS supporters on Twitter. *The Brookings Institution.*

Bryden, J., Funk, S., Geard, N., Bullock, S., Jansen, V.A.A., 2011. Stability in flux: community structure in dynamic networks. *Journal of the Royal Society Interface*. 8(60):1031-40.

Bryden, J., Funk, S., Jansen, V.A.A., 2013. Word usage mirrors community structure in the online social network Twitter. *EPJ Data Science*. 2(3)

Chatfield, A.T., Reddick, C.G., Brajawidagda, U., 2015. Tweeting propaganda, radicalization and recruitment: Islamic state supporters multi-sided twitter networks. *Proc 16th Annual International Conference on Digital Government Research*. pp.239-49 DOI:10.1145/2757401.2757408

Fisher, A., 2015. Swarmcast: How Jihadist Networks Maintain a Persistent Online Presence. *Perspectives on terrorism*. 9(3)

Gladstone, R., 2015. Twitter Says It Suspended 10,000 ISIS-Linked Accounts in One Day. *The New York Times*. April 9.

Goodman, L.A., 1961. Snowball Sampling. Ann Math Stat 32:148–170.

Greene, K.J., 2015. ISIS: Trends in Terrorist Media and Propaganda. *International Studies Capstone Research Papers.* Paper 3.

Leavitt, A., 2014. From #FollowFriday to YOLO: Exploring the Cultural Salience of Twitter Memes. In: Weller, Bruns, Burgess, Mahrt, Puschmann. ed. *Twitter and Society*. New York: Peter Lang. pp.137-154.

Levy, R., 2014. ISIS Tries to Outwit Social Networks. *Vocativ*. June 17

Magdy, W., Darwish, K., Weber, I., 2015. #FailedRevolutions: Using Twitter to Study the Antecedents of ISIS Support. arXiv:1503.02401v1

Mahmood, S., 2012. Online social networks: The overt and covert communication channels for terrorists and beyond. *2012 IEEE Conference on Technologies for Homeland Security (HST).*

McPherson, M., Smith-Lovin, L., Cook, J.M., 2001. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*. 27:415-44.

Miller, G.A., 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*. 63(2):81–97.

Moriarty, B., 2015. Defeating ISIS on Twitter. *Technology Science*. 2015092904

Ryan, L., 2014. Al-Qaida and ISIS Use Twitter Differently. Here's How and Why. *National Journal.* October 9.

Standing, L., 1973. Learning 10,000 pictures. *Q J Exp Psychol.* 25:207-22.

Stern, J., Berger, J.M., 2015. *ISIS: The state of terror.* London: William Collins

Tamburrini, N., Cinnirella, M., Jansen, V.A.A., Bryden, J., 2015. Twitter users change word usage according to conversation-partner social identity. *Social Networks.* 40:84-89.

Weimann, G., 2014. *New Terrorism and New Media*. Washington, DC: Commons Lab of the Woodrow Wilson International Center for Scholars.

Supplementary Material

Why a quantitative approach?

Manually inspecting the complete dataset of 1,920 users for replicates would be very time consuming. Berger's sample of 46,000+ would make the task close to impossible. The feasibility of this task is partly limited by its reliance on human memory capacity. Whilst working memory capacity is a mere 7±2 items (Miller, 1956), we suggest that a more appropriate indicator is recognition memory – the ability to recognise whether or not something matching the account had been encountered earlier in the dataset. Standing (1973) empirically derived equations showing that recognition memory follows a power law with the number of items presented. We can therefore calculate that if humans inspected our 1,920 accounts as printed words, Standing's work predicts the number capable of being held in memory is:

10^((0.92*log(1,920 items))-0.01) = 1,025

Since for many accounts we also have a screenshot of their Twitter profile, Standing's equation for pictorially presented data predicts:

10^((0.93*log(1,920 items))+0.08) = 1,360

The upper limit of human memory whilst attempting a match search with our medium sized dataset is therefore ~53-71% of previously encountered accounts. Since each account is actually represented by around 10 words, not one, this oversimplification generates an extremely conservative upper limit. Re-calculating for Berger's conservative estimate of 46,000 ISIS accounts, only ~41-56% can be held in recognition memory; another overestimation. Standing's results may also not generalise this far beyond the 10,000 items used in his work.

To aid the quick finding of resurgent accounts, we therefore used a quantitative approach to draw our attention to several accounts at a time. Hypothetically, the simplest approach might have been selecting two random accounts to evaluate simultaneously. This would have been ineffective. A quantitative approach should work on an assumption or hypothesis about the data. We assumed that finding matches would be aided by selecting accounts whose biographies, names and locations contained >30% of the same words. This meant that only accounts with these attributes had sufficient information to analyse.

6. Evaluating Machine and Crowdsourcing Methods for Classifying Pseudoreplicate Terrorist Supporting Accounts on Twitter

Overview in relation to the thesis

The previous chapter demonstrated how problematic resurgent terrorist accounts are and thus the importance of finding and controlling for them. Studying the landscape of twitter accounts is therefore important to get a better picture of how they work. The evidence from the literature review and previous chapter give both a good understanding of how to identify terrorist-linked accounts and suggest that the Baqiya family is a good option for study. The question addressed, therefore, is how to generate a reliable set of Twitter accounts.

This research chapter addresses whether automated machine methods can improve our ability to reliably find resurgents. A text similarity based machine model, based on the similarity heuristic used in the previous chapter, is developed and validated relative to against human-annotation. Although it follows the machine learning and evaluation approach, it overcomes problems unique to social media and terrorism studies, whereby objectively validated learning and test datasets are challenging to come by. This work contributes novel, validated methods for finding resurgents, whereas previous work has relied on human annotation and case studies on only one (Stern and Berger, 2015) or

four (Berger and Perez, 2016) users, or else has been unable to find and control for resurgents at all (Berger and Morgan, 2015; Chatfield *et al.*, 2015; Magdy *et al.*, 2015).

Wright, S., Denney, D., Pinkerton, A., Jansen, V.A.A., Bryden, J., 2016. Evaluating Machine and Crowdsourcing Methods for Classifying Pseudoreplicate Terrorist Supporting Accounts on Twitter. [*In preparation*]

Evaluating Machine and Crowdsourcing Methods for Classifying Pseudoreplicate Terrorist Supporting Accounts on Twitter

Shaun P Wright[1,2,3*], Alasdair Pinkerton[1], David Denney[2], Vincent AA Jansen[3], John Bryden[3]

[1] Department of Geography, Royal Holloway University of London, Egham, Surrey, United Kingdom

[2] School of Law, Royal Holloway University of London, Egham, Surrey, United Kingdom

[3] School of Biological Sciences, Royal Holloway University of London, Egham, Surrey, United Kingdom

* Corresponding author

E-mail: shaun.wright.2013@live.rhul.ac.uk (SPW)

Abstract

While social media websites often ban users who have malicious or illegal intentions, these users often simply recreate their accounts. Sampling over time, this problem can create skewed data sets with duplicates of users who are difficult to rigorously analyse. Identifying such duplicates can be overly time consuming for humans, however automated approaches, while faster, are not guaranteed to be trustworthy. Here, we develop and evaluate a text-similarity based model. We also show that annotators from

different backgrounds and expertise levels agree and make their decisions using features that are just as amenable to machine methods.

1 Introduction

People often have multiple, online accounts (Korula and Lattanzi, 2014; Malhotra, 2013; Vesdapunt and Garcia-Molina, 2014; Goga *et al.*, 2015). Law enforcement, marketing and sociological or political analysis could all benefit from methods that easily match these accounts together. This paper attempts to develop and validate machine methods to do just that; in particular in the context of matching multiple, extremism- or terrorism-supporting accounts that belong to a single individual.

Terrorists and their supporters are very active on Twitter (Weimann, 2014; Berger and Morgan, 2015). As of 2015/16, just as offline, amongst terrorism supporter, Daesh was predominant (Berger and Morgan, 2015). On Twitter, these supporters form "a loose network of Islamic State supporters from around the world who share news, develop close friendships, and help each other" (Amarasingam, 2015)—'the Baqiya family'. Although repeatedly suspended by Twitter, Baqiya family members can easily create new accounts and resurge back (Wright *et al.*, 2016; Stern and Berger, 2015; Greenberg, 2015).

Thus, with tens of thousands of Baqiya family accounts active every month (Berger and Morgan, 2015), analysts of terrorism and intelligence struggle to find and control for resurgents (Wright *et al.*, 2016; Berger and Morgan, 2015; Berger and Perez, 2016;

Stern and Berger, 2015; Chatfield *et al.*, 2015; Magdy *et al.*, 2015). Sampling large numbers of accounts under the false assumption that each account represents a unique individual can, therefore, lead to datasets that are biased by pseudoreplicates (Hurlbert, 1984; Vaux *et al.*, 2012; Wright *et al.*, 2016)—replicates that are not statistically independent. Such pseudoreplication can also occur in other areas of research—especially where multiple accounts exist for malicious purposes, such as bullying, grooming, or spam generation. In this paper, we develop and validate a novel machine method to tackle the important social media classification problem that is identifying and matching accounts—amongst the large and noisy volume of Twitter data—that are multiple instances of the same terrorism-supporting users.

Similar work identified multiple, "fake" accounts belonging to the same users in the Twitter population as a whole (Gurajala, 2015). Gurajala showed that machine filtering of identical profile features and near identical handles can reliably classify pseudoreplicates (Gurajala, 2015; Brynielsson *et al.*, 2012). Just as with human-annotation (Harris and Srinivasan, 2014), however, machine methods may suffer domain-dependent performance differences. This is especially true within the terrorism domain, where identities are concealed and contacting users to verify matches is unrealistic. Thus, this study also investigates whether machine approaches reliably identify resurgent terrorist supporters within the terrorism domain.

Our novel machine method uses the salient information contained in publicly available profile features—such as names, @handles, locations and biographies. Such features have previously enabled the matching of users who exist across multiple social

networks (Korula and Lattanzi, 2014; Malhotra, 2013; Vesdapunt and Garcia-Molina, 2014; Goga *et al.*, 2015). Furthermore, within Twitter, public profile features have enabled machine methods to classify accounts and predict characteristics such as income (Preoțiuc-Pietro *et al.*, 2015) or marketing interests (Lo *et al.*, 2015). Our new method uses publicly available profile features to solve a functionally similar matching problem: pseudoreplicates that exist across multiple, longitudinal, Twitter samples. Although we (Wright *et al.*, 2016) and others (Stern and Berger, 2015) have observed that matching, terrorist supporting accounts accounts have near identical handles (Gurajala, 2015), we have further observed that accounts are often nearly identical in other metadata fields: name, biography and location. Consequently, filtering out non-identical names, locations and biographies can generate false negatives—less problematic for Gurajala (2015), whose aim was to find sufficient pseudoreplicates to characterise further and focused on avoiding false positives as a cross-section of the whole of Twitter returned ~56,000 pseudoreplicates. Our aim in this study, however, is to find all of the resurgents within our sample; minimising both false negatives and false positives. Thus, we extend the method to allow all four metadata fields to be nearly identical (as measured with the Bray-Curtis similarity index (Bray and Curtis, 1957; see methods)). In this study, we evaluate our novel method, based upon publicly available profile features, for classifying and predicting pseudoreplicates within Twitter.

We adopt the standard machine-learning-evaluation approach—comparison of the method's classifications against a test-dataset of known, pre-characterised data (Table 1). With this approach, although the pre-characterised dataset is assumed to represent the 'truth' and be a 'gold standard' to be replicated (Cormack and Lynam, 2005; Yang and Srinivasan, 2014), fitted models are usually accepted to be only approximations—

albeit useful approximations with predictive validity. In social science and studies of social networks, however, datasets is often ambiguous or subjective, making human-annotation the primary way to acquire pre-characterised, training and test-datasets (Cormack and Lynam, 2005; Harris and Srinivasan, 2014; Yang and Srinivasan, 2014; Lo *et al.*, 2015). Irrespective of how methodically annotated, the human-annotated dataset against which models are fitted will also be an approximation. The fact that the "model and the data are two moving targets that we try to overlay one upon the other" (Rykiel, 1996) means that "we cannot assume that data accurately represent the real system and therefore constitute the best test of the model". As human-annotation is currently the only standard method for finding pseudoreplicate terrorist accounts, however, we argue that validating our new machine alternative against it is an important step. The first part of our study evaluates the performance of our new machine method, based on publicly available profile features, against the main method in use at the moment—human-annotation.

Table 1. The machine learning performance evaluation approach.

		Machine method/model output	
		Pseudoreplicate match	Not pseudoreplicate match
Test-dataset	match	True Positive (TP)	False Negative (FN)
	no match	False Positive (FP)	True Negative (TN)

Performance is evaluated by categorising each of a method or model's positive outputs (pseudoreplicate) as a True Positive or False Positive and each of its negative outputs (not a pseudoreplicate) as a True Negative or False Negative, by comparing them with what the known, pre-characterised, test-dataset says that they should be.

The human-annotation was done—as is outlined further in the methods—by the research team and a limited number of undergraduates and other academics at Royal Holloway University of London. While, in other domains, human classification has been improved through crowdsourcing (Cormack and Lynam, 2005; Smucker *et al.*, 2012), the ethics of using crowdsourcing platforms (for example Amazon Mechanical Turk) for extremist and terrorist content are more challenging and thus we were limited to volunteers in the controlled, university environment and university ethical approval.

Given that limitation, in the second part of this study we address the concern that human-annotation may not provide a very accurate or high-quality test-dataset. With the main approach for finding and matching pseudoreplicates—at present, manual inspection and classification of data by humans—only case studies of terrorist-supporters have emerged—and only on five such extremists (Stern and Berger, 2015; Berger and Perez, 2016). We hypothesise that this is due to the limitations of human memory when working with datasets in the tens of thousands (Wright *et al.*, 2016, Supplementary Material). On the one hand, our novel machine method should improve on human classification, just like crowdsourcing (Cormack and Lynam, 2005; Smucker *et al.*, 2012), machine method (Lo, 2015), or a mixture of the two (Harris and Srinivasan, 2014), approaches have in other domains—the Text Retrieval Conference (TREC) crowdsourcing track (Cormack and Lynam, 2005; Smucker *et al.*, 2012), which invited attempts to develop crowdsourcing (later widened to include machine) methods to emulate human-classification of documents is an example of one such effort. On the other hand, the same arguments for replacing human-annotation with machine methods —human-annotation being temporally, mentally and financially expensive—lead to a

hypothesis that human-annotation does not generate particularly high-quality test-datasets.

We demonstrate this this is the case, aided by a quirk in the data sampled. During sampling, Twitter suspended some accounts that we had already sampled. As we moved suspended accounts to a separate “inactive dataset”, where we recorded their metadata but they no longer contributed to further sampling, some accounts that subsequently had their suspension lifted by Twitter were re-sampled into the “active dataset”. Where multiple, sampled account share a unique Twitter ID, they can be objectively identified as 'positive controls'—multiple instances by the same author. As these were discovered during analysis, after human-annotation of the dataset, we therefore benefited from the unplanned ability to evaluate human-annotation against 'positive controls'. They were not used further, however, as they were not included in the experimental design and the machine methods—by loading accounts by Twitter ID—were explicitly coded in such a way as to never compare two 'positive controls'; the implications and future opportunities are considered further in the discussion section.

Given the demonstrable limitations in the existing method of generating the human-annotated, test-dataset, our hypothesis that the machine method will perform better causes circular problems for evaluation. Combined with our limited access to crowdsourcing, we designed an alternative protocol to validate the machine method. The poor performance of human-annotation is, in part, because much of the humans' time is spent evaluating negative, non-matches in the dataset (Yang and Srinivasan, 2014). This annotation is, in effect, attempting to remember a large dataset and recalling whether an account has been previously encountered—something at which humans are

poor (Wright *et al.*, 2016, Supplementary Material). Humans are, however, better at visual comparison of a limited number of items and identifying whether they match. In order to do a more rigorous evaluation of our new machine method, in the second part of this study, we therefore get humans to individually annotate, *post hoc*, the outputs of our machine method (along with some negative controls).

The final section of our study attempts to determine how human-annotators are making their decisions, so that future machine methods might be able to take these into account. We record the reasons given by annotators and test whether expertise (Harris and Srinivasan, 2014) and features such as pictorial information are important. Showing that our machine method assesses the data in the same way as human-annotators from a range of expertise levels would further support the validity of our feature based approach and suggest that the expense of expert human annotation can be reduced for some domains and research questions.

In this paper, therefore, we first develop a novel machine approach, based on the similarity of publicly available features, and evaluate it against a human-annotated, test-dataset. After we then demonstrate the inaccuracy of human-annotation by evaluating it against 'positive controls', we re-evaluate our novel machine method against *post hoc* human-annotation of its outputs, showing that this improves the assessment of its performance. Finally, we test the importance of expertise and other features to human-annotators.

2 Materials and Methods

2.1 Dataset

We sampled 1,920 jihadist, jihadist supporting, or jihadist linked Twitter accounts; the same dataset sampled and characterised in Wright *et al.* (2016). We snowball sampled (Goodman, 1961) between May and July 2015 using the Twitter API. We seeded the sample with 34 English-speaking jihadist accounts identified through manual analysis of Twitter, aided by Twitter's "Who to follow" suggestions. To build a highly intra-linked jihadist Twitter community, we weighted the snowball sampling. This is based on the principle of homophily: the tendency of people to associate with similar people (McPherson *et al.*, 2001) and bias their interactions to members of the same community with whom they share a social identity (Bryden *et al.*, 2011; 2013; Tamburrini *et al.*, 2015). We therefore added, daily, any account followed by >10% of the users already in our sample and with <1,000 followers of its own. Further description of this methodology and rationale is available in Wright *et al.* (2016). Following the end of metadata sampling, we took a screenshot of each Twitter account in our sample. 602 of the accounts had been suspended and 97 no longer existed due to name changes. For those accounts, we obtained only a screenshot of Twitter's notification.

2.2 Machine method

2.2.1 Bray-Curtis similarity model

Our machine model, based upon the previously identified heuristic that resurgents have near similar profile features (Wright *et al.*, 2016), uses the Bray-Curtis similarity index (Bray and Curtis, 1957) to compare two strings of text. We calculate the proportion of words shared between texts, where each instance of a word is treated independently. For each account, we appended the lower-case biography, name, handle and location into one metadata string. Then for each pairwise combination of the 1,920 accounts, we calculated the similarity of their metadata strings. We then produced two models by varying the threshold level of similarity required to predict that two accounts formed part of the same resurgent set: Bray-Curtis$_{30}$—accounts >30% similar; Bray-Curtis$_{50}$—accounts >50% similar. We applied this rule transitively to build sets containing all pseudoreplicate instances of a resurgent (so if A=B and B=C, then A=C; even if the model does not give A=C).

2.2.2 Control models

To demonstrate that our algorithm was tapping into meaningful information rather than matching accounts by chance, we used two negative control model ('negative' controls given that we hypothesise the matches within are not real matches).

'Random' control model

For every set of accounts matched by the Bray-Curtis$_{30}$ model, we generated a neutral, random, negative control prediction of the same size by randomly selecting users,

without replacement, from the entire sample of users. We repeated this with the Bray-Curtis$_{50}$ model to create two random controls, controlled for the distribution of set sizes (RC_{30} and RC_{50}).

'Alphabetical' control model

As resurgent accounts use similar or numerically increasing handles (Stern and Berger, 2015), sorting the sample alphabetically by handle would be a naïve, but more relevant, negative control than random selection. Therefore, after sorting the sample, for every set of resurgents matched by the Bray-Curtis$_{50}$ model, we generated an alphabetical, negative control set of the same size by grouping consecutive accounts, with a randomly selected starting point. If this ran over the end of the set then we continued from the beginning. Accounts were able to feature in more than one set. Again, we repeated this for every set of resurgents predicted by the Bray-Curtis$_{30}$ model to create two alphabetical controls, controlled for the distribution of set sizes (AC_{30} and AC_{50}).

2.3 Evaluating performance

2.3.1 Evaluating performance—machine method against human-annotation

For each pairwise combination of the 1,920 sampled accounts, we compared whether they were classified as a match by each model (Bray-Curtis; Neutral Control; Alphabetic Control) against whether they had been classified as such in the human-annotated, test-dataset. We then evaluated performance using each of the standard machine learning metrics given in Table 2 of Section 2.4—'Performance statistics':

True Positive—machine method **matched** / human-annotated test-dataset **matched**;

False Positive—machine method **matched** / human-annotated test-dataset **did not match**;

True Negative—machine method **did not match** / human-annotated test-dataset **did not match**;

False Negative—machine method **did not match** / human-annotated test-dataset **matched**;

Human-annotation of the accounts was done manually by the authors (omitting the lead author who, having sampled and prepared the data, had learnt many of the resurgents for which to search). We inspected accounts on a computer screen, in a folder containing a pdf document for each of the 2,144 accounts (including positive control accounts). Each pdf document contained one account, shown as a uniform table containing handle, name, location and biography, along with a screenshot of the twitter profile. As no researcher had time to inspect more than a subset of the entire dataset and this was a test of current best practise, we informed each researcher of the accounts that had already been inspected and those that had been matched.

Our initial experimental design considered an inter-coder agreement criteria, for example whereby two or more annotators, by majority or unanimity, would have to match a pair of account for it to be accepted for analysis. The time constraints—and lack of access to online crowdsourcing, as mentioned earlier—meant we were limited in human-resources and no annotator viewed the same subset of 2,144 accounts as any other user; i.e. we priorities breadth of coverage over depth. Thus, if any annotator matched accounts, we accepted that match for analysis. We did, however, provide

annotators with the outputs of previous annotators and they added additional accounts to the previously identified sets of matched accounts; they did not remove any previously identified matches.

Further, given that our aim in the final section of our study was to investigate 'how' human-annotators make their decisions—using which features—we decided to prescribe no *a priori* coding scheme. We provided annotators with a form to note down brief reasons for each classification decision.

In the instructions, we also provided annotators with some Islamic and Arabic terms to prevent terms such as “abu” (meaning “father of”) being misinterpreted as names and thus be given too much salience as an indicator that two accounts belonged to the same user (whilst also pointing out that the word provided extra gender information).

Three members of the research team analysed 2,144 randomly shuffled accounts. In two hours each, six hours total, they managed to inspect 957 (44.6%) of the accounts between them.

2.3.2 Evaluating performance—human-annotation against positive controls

As discussed in the introduction, we hypothesise that human-annotation may not provide a high-quality test-dataset—due to limitations on human memory when remembering large datasets (Wright *et al.*, 2016, Supplementary Material). The standard machine learning approach uses objectively-known, test-data ('positive controls') to evaluate a model. As human-annotation is the usual method of generating test-datasets

for social data, it is circular to use a human-annotated, test-dataset to evaluate human-annotation.

The 'positive controls' used instead are a quirk of the data sampling. During sampling, Twitter suspended some accounts that we had already sampled. As we moved suspended accounts to a separate “inactive dataset”, where we recorded their metadata but they no longer contributed to further sampling, some accounts that subsequently had their suspension lifted by Twitter were re-sampled into the “active dataset”. Where multiple, sampled account share a unique Twitter ID, they can be objectively identified as 'positive controls'—multiple instances by the same author. There are 118 account duplicated once in this way and one account duplicated twice; a total of 121 positive control pairs.

For each pairwise combination of the 2,144 sampled accounts, we compared whether they were classified as a match in the human-annotated, test-dataset and in the 'positive control' test-dataset. We then evaluated performance using each of the standard performance statistics:

True Positive—human-annotation test-dataset **matched** / positive controls **matched**;
False Positive—human-annotation test-dataset **matched** / positive controls **did not match**;
True Negative—human-annotation test-dataset **did not match** / positive controls **did not match**;
False Negative—human-annotation test-dataset **did not match** / positive controls **matched**;

2.3.3 Evaluating performance—machine method against *post hoc* human-annotation

After demonstrating the poor quality of a human-annotated, test-dataset that required the annotator to remember the entire dataset, we used a protocol that did not require human memory, rather their ability to inspect and compare a limited number of model outputs.

We analysed all positive outputs from the Bray-Curtis models, after meeting our ethics requirements and excluding outputs containing graphic or distressing images (15%) (Bray-Curtis$_{30}$ model, n=156; Bray-Curtis$_{50}$ model, n=146). Due to the limited availability of human-annotators, we randomly selected a subset of outputs from each negative control (n=50) and alphabetical control (n=100) model.

We randomly shuffled and anonymised outputs. We formatted outputs in pdfs as before, although all accounts matched within a single output were included in the same pdf. Annotators were given an on-screen folder containing a pdf for each output they were to annotate.

As before, annotators used (and recorded) their own criteria to decide whether each output was a: full match—all accounts within the pdf matched all other accounts; part match—some accounts matched, but at least one did not; or no match—none of the accounts in the pdf matched any of the others. As an inter-coder criteria, we took the majority decision in the event of disagreement between multiple annotators (if this was tied, we reverted to no match).

We evaluated performance over each pairwise combination of the 1,920 sampled accounts using each of the standard performance statistics (recording part-match constituent pairs as did not match):

True Positive—machine method **matched** / *post hoc*-annotation **matched**;
False Positive—machine method **matched** / *post hoc*-annotation **did not match**;
True Negative—machine method **did not match** / *post hoc*-annotation **did not match**;
False Negative—machine method **did not match** / *post hoc*-annotation **matched**;

We also combined the *post hoc* annotation, test-dataset with the human-annotated, test-dataset to get an improved evaluation:
True Positive—machine method **matched** / either test-dataset **matched**;
False Positive—machine method **matched** / both test-datasets **did not match**;
True Negative—machine method **did not match** / both test-datasets **did not match**;
False Negative—machine method **did not match** / either test-dataset **matched**;

2.4 Performance statistics

We adopt the standard machine learning evaluation approach and corresponding performance metrics (accuracy, precision, recall/sensitivity, specificity, G mean, F score and LEM) (Table 2), noting documented flaws (Cormack and Lynam, 2005; Powers, 2011; Sokolova *et al.*, 2006). After calculating the numbers of True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN) produced by each model, we calculated each performance statistic.

Table 2. Performance statistics

Statistic	Formula	Summary
TPR / sensitivity / recall	TP / (TP + FN)	Proportion of the test-dataset pseudoreplicate matches found by the model.
TNR / specificity	TN / (TN + FP)	Proportion of the test-dataset negatives (not pseudoreplicate matches) correctly identified as such by the model.
Accuracy	(TP + TN) / (TP+TN+FP+FN)	Proportion of all test-dataset correctly classified by the model.
Precision / Positive Predictive Value	TP / (TP + FP)	Proportion or confidence that pseudoreplicates identified by the model are relevant and correct (i.e. not false positives).
G mean	Sqrt(TPR*TNR)	The geometric mean of precision and recall. Normalizes the true pseudoreplicates found by the model against the geometric mean of the number of pseudoreplicates found and the number in the test-dataset (Powers, 2011).
F score	2* (Precision * Recall)/(Precision + Recall)	The harmonic mean of precision and recall. References the true pseudoreplicates found by the model against the arithmetic mean of the number of pseudoreplicates found and the number in the test-dataset (Powers, 2011).
LAM	logit^{-1} (½* (logit(FNR) + logit(FPR)); logit(x) = log(x/(1–x)); $\text{logit}^{-1}(x) = e^x/(1+e^x)$	"This measure imposes no *a priori* relative importance on [positive or negative] misclassification and rewards equally a fixed-factor improvement in the odds of either" (Cormack and Lynam, 2005) (Smucker *et al.*,

		2012)

The names, formulas and descriptions of 'what they do' and 'why they are used' of standard summary statistics for evaluating the performance of machine methods (Cormack and Lynam, 2005; Powers, 2011; Sokolova *et al.*, 2006).

2.5 Assessing features of human-annotation

2.5.1 Expertise

To demonstrate that neither annotation, nor the improved human annotation test dataset, is dependent on subjective opinion, specialist expert knowledge or academic training, we used three categories of model annotators: core research team, academics external to the core research team and undergraduates. We correlated the classifications given by combinations, pairwise, of annotator types.

Research team markers—again, we excluded the lead author whose overexposure to the data in downloading and preparing meant many of the resurgents to search for were known. Each member of the core research team (n=3) marked 120 or 121 resurgent sets each. External academic markers—(n=2) of at least post-doctoral standing marked 60 resurgent sets each. Undergraduate markers—with prior (though not expert) knowledge of jihadist terrorism via taking one or two term, third year courses on Risk, Insecurity and Terrorism. Each undergraduate (n=5) marked 60 or 61 resurgent sets each.

Using the outputs that were annotated by users from multiple expertise levels, annotated with ordinal “no”, “part” or “full” match, we used Spearman's rank correlation coefficient to correlate each expertise group with each of the others: research team with

external academics; research team with undergraduates; and external academics with undergraduates, with a Bonferroni correction factor of three.

2.5.2 Public profile features

During every human annotation, brief reasons for each classification decision were recorded by the marker. Inspection of the repeated reasons (sometimes recorded as "ditto") revealed clear categories: name, handle, biography, imagery, numbering of the account, location, completely identical, key word, gender and other information given in the content. No reasons fell outside these categories.

For each type of annotator (whole dataset annotation, research team, external academic, undergraduate), we calculated the proportion of reasons given that fell into each category. We then used a chi-square test (categorical data) to test the null hypothesis that reasons given were independent of expertise level (df = 18).

2.5.3 Images

We investigated further the importance of imagery to humans when annotating the datasets. Following the end of metadata sampling, we took a screenshot of each Twitter account in our sample. 602 of the accounts had been suspended and 97 no longer existed due to name changes. For those accounts, we obtained only a screenshot of Twitter's notification. Where accounts had screenshots, these were included in the pdf documents presented to annotators. We compared the proportion of screenshots found by each of the models with the proportion in the entire dataset (28.5%) (one-sample t-test).

2.6 Ethics statement

The methods used in this study were approved by the Royal Holloway University of London University Research Ethics Committee (REC ProjectID: 15).

3 Results

3.1 Evaluating performance—machine method against human-annotation

We first developed and evaluated our new machine method. We applied the Bray-Curtis text-similarity based machine model (BC_{50}), the Neutral / Random Control model (RC_{50}) and the Alphabetic Control model (AC_{50}) to the 1,920 unique Twitter accounts in the sample in order to identify pseudoreplicates. The performance of the three models, as evaluated against the test dataset of human annotation of the entire dataset, is shown in Table 3.

Human-annotation of the 2,144 accounts (including positive controls) identified 40 pairs of matching, pseudoreplicate accounts. Grouping the overlapping pairs into larger sets revealed 20 sets of matching pseudoreplicates (i.e. 20 unique resurgents). Each set contains an average of 2.4 accounts (mean ± std. 0.821), or 48 accounts in total.

Treating human-annotation as the test-dataset, our Bray-$Curtis_{50}$ model seemingly missed the pseudoreplicates (recall = 20.0%) and generated many false positives (precision = 0.481%). Although both controls missed all the pseudoreplicates (both

recall = 0%). The Bray-Curtis$_{30}$ model performed almost identically to the Bray-Curtis$_{50}$ model (S1 Table).

Table 3. Machine method performance against human-annotated, test-dataset: BC_{50}, RC_{50} and AC_{50}.

Machine model:	**RC_{50}**	**AC_{50}**	**BC_{50}**
True Positives	0	0	8
True Negatives	2,045,530	2,045,762	2,045,581
False Positives	1706	1474	1655
False Negatives	40	40	32
TPR / sensitivity / recall	0%	0%	20.0%
TNR / specificity	99.9%	99.9%	99.9%
Accuracy	99.9%	99.9%	99.9%
Precision	0%	0%	0.481%
G mean	0.00	0.00	0.447
F score	0.00	0.00	0.00939
LAM	#DIV/0!	#DIV/0!	0.224

Performance of three models evaluated against the human-annotated, test-dataset: Neutral / Random Control (50%), Alphabetic Control (50%) and Bray-Curtis (50%). All results given to 3.s.f.

As an aside, we can get an indication of how humans decided on their annotations. Since the dataset of 2,144 accounts were presented on screen, in a folder containing a pdf document per account, we can look at the relationship between how far apart two accounts were in the folder and how likely the annotators were to match them. Of the 121 pairs of positive control accounts, 20 were within 250 documents of one another. Human annotation identified 12 of these (60% recall / sensitivity). The remaining 101 pairs of positive controls were more than 250 documents apart and human annotation

identified none (0% recall / sensitivity) of these. This supports the hypothesis that human annotation of the entire dataset is heavily hampered by human memory limitations.

3.2 Evaluating performance—human-annotation against positive controls

There were 121 pairs of accounts that had been reinstated by Twitter and duplicated in the sample. These formed positive controls that we can objectively call pseudoreplicates. Human-annotation only identified 12 (9.91%) of these (Table 4). Table 4 shows how well human-annotation of the entire dataset performs at finding pseudoreplicates if the positive controls are used as the test-dataset and the other (non positive control) accounts identified by human-annotation are treated as false positives.

Table 4. Performance statistics for human annotation of the entire dataset.

Test-dataset:	Positive Controls
True Positives	12
True Negatives	2,297,135
False Positives	40
False Negatives	109
TPR / sensitivity / recall	9.91%
TNR / specificity	100%
Accuracy	100%
Precision	23.1%
G mean	0.315
F score	0.139
LAM	0.130

Statistics evaluating the performance of human-annotation at finding pseudoreplicates in the entire dataset. Human-annotation is evaluated against the positive control test-datasets. All results given to 3.s.f.

3.3 Evaluating performance—machine method against *post hoc* human-annotation

Having demonstrated the limitations in the existing method of generating the human-annotated, test-dataset, we therefore had humans individually annotate, *post hoc*, the outputs of our machine method (along with some negative controls) (Table 5). This time, precision increased to 47.6%, indicating that many of the model outputs evaluated against the human-annotated, test-dataset as False Positives were not actually so. The number of found pseudoreplicates has increased and the number of missed pseudoreplicates now constitutes a smaller proportion of the overall data (recall = 95.5%), due to many of the pseudoreplicates found by the model now being annotated

as True Positives. Fig 1 shows a graphical representation of the tendency of the human-annotators to classify, *post hoc*, the Bray-Curtis$_{50}$ output as True Positive "matches" (each account in the output matched each other account), but the control model outputs as False Positives—"part-matches" (some accounts in the output matched, but at least one did not) or "no-matches" (not a single account in the output matched any other).

Table 5. Machine method performance against human-annotated, test-dataset and *post hoc*-annotation: BC_{50}, RC_{50} and AC_{50}.

	Human-annotated, test-dataset (*Table 3 repeated*)			***Post hoc* human-annotation**		
Machine model:	RC_{50}	AC_{50}	BC_{50}	**RC_{50}**	**AC_{50}**	**BC_{50}**
True Positives	0	0	8	0	62	672
True Negatives	2,045,530	2,045,762	2,045,581	2,046,770	2,046,397	2,045,831
False Positives	1706	1474	1655	446	777	740
False Negatives	40	40	32	40	40	32
TPR / sensitivity / recall	0%	0%	20.0%	0%	60.8%	95.5%
TNR / specificity	99.9%	99.9%	99.9%	100%	100%	100%
Accuracy	99.9%	99.9%	99.9%	100%	100%	100%
Precision	0%	0%	0.481%	0%	7.39%	47.6%
G mean	0.00	0.00	0.447	0.00	0.212	0.977
F score	0.00	0.00	0.00939	0.00	0.132	0.636
LAM	#DIV/0!	#DIV/0!	0.224	#DIV/0!	0.141	0.0845

Performance of three models evaluated against the human-annotated, test-dataset and *post hoc* human-annotation: Neutral / Random Control (50%), Alphabetic Control (50%) and Bray-Curtis (50%). All results given to 3.s.f.

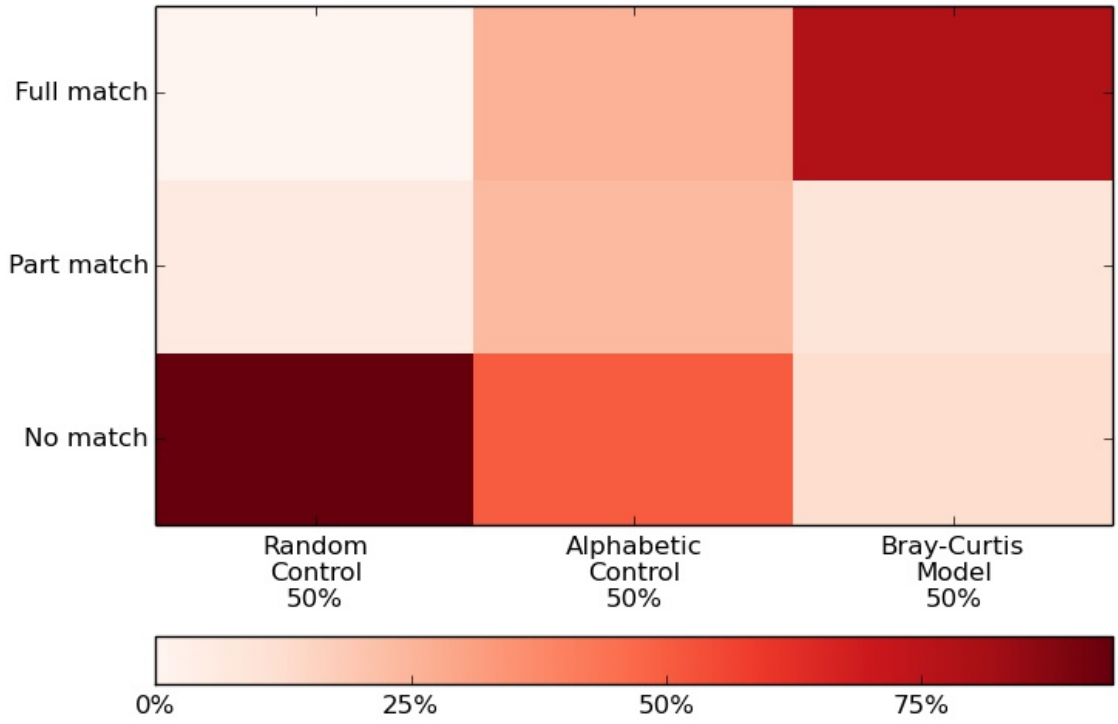


Fig 1. Text-similarity based model produces true positives whilst control models produce false positives. Heatmap showing the judgements made about the sets of pseudoreplicate accounts matched together by three models, as evaluated by improved human annotation. Annotators tended to classify, *post hoc*, the Bray-Curtis$_{50}$ output as True Positive "matches" (each account in the output matched each other account), but the control model outputs as False Positives—"part-matches" (some accounts in the output matched, but at least one did not) or "no-matches" (not a single account in the output matched any other). The darker the red, the greater the proportion of that model's outputs fell into that category. 93.8% of the predictions made by the random control model and 50.3% of the alphabetic control model, were judged as no match, whereas 78.4% of the predictions made by the Bray-Curtis model were evaluated as full match.

Evaluated against both test datasets, all models and control models have accuracy > 99.9% (Table 5) and appear to perform equally. This, however, is because they are

equally good at correctly identifying true negatives (all specificity > 99.9%), of which the test datasets are over 99.9% composed. This bias is why the G mean and F score (Table 5) are better measures of performance and clearly show the improved performance of the Bray-Curtis$_{50}$ model relative to the control models.

The Bray-Curtis$_{30}$ model (S1 Table) performed almost identically to the Bray-Curtis$_{50}$ model, although the lower threshold meant that more false positives were included and thus the precision was slightly lower (16.7%—improved human annotation). A graphical comparison can also be found in the Supporting Information (S1 Fig).

3.4 Assessing features of human-annotation

3.4.1 Expertise

We tested the similarity of annotation between annotators from three different expertise levels and backgrounds (research team, other academics external to the research team and undergraduates) and demonstrated significant correlations (Table 6) between each pair. The similarity is shown graphically in Fig 2.

Table 6. Correlations between the classifications given by the different categories of annotator.

Correlated pair	n	r^2	p
Research team and external academic markers	90	0.601	2.95×10^{-19}
External academic and undergraduate markers	90	0.551	5.51×10^{-17}
Research team and undergraduate markers	296	0.663	1.84×10^{-71}

Spearman's ranked correlations between the "no-", "part-" and "full-" match classifications given by the different categories of annotator.

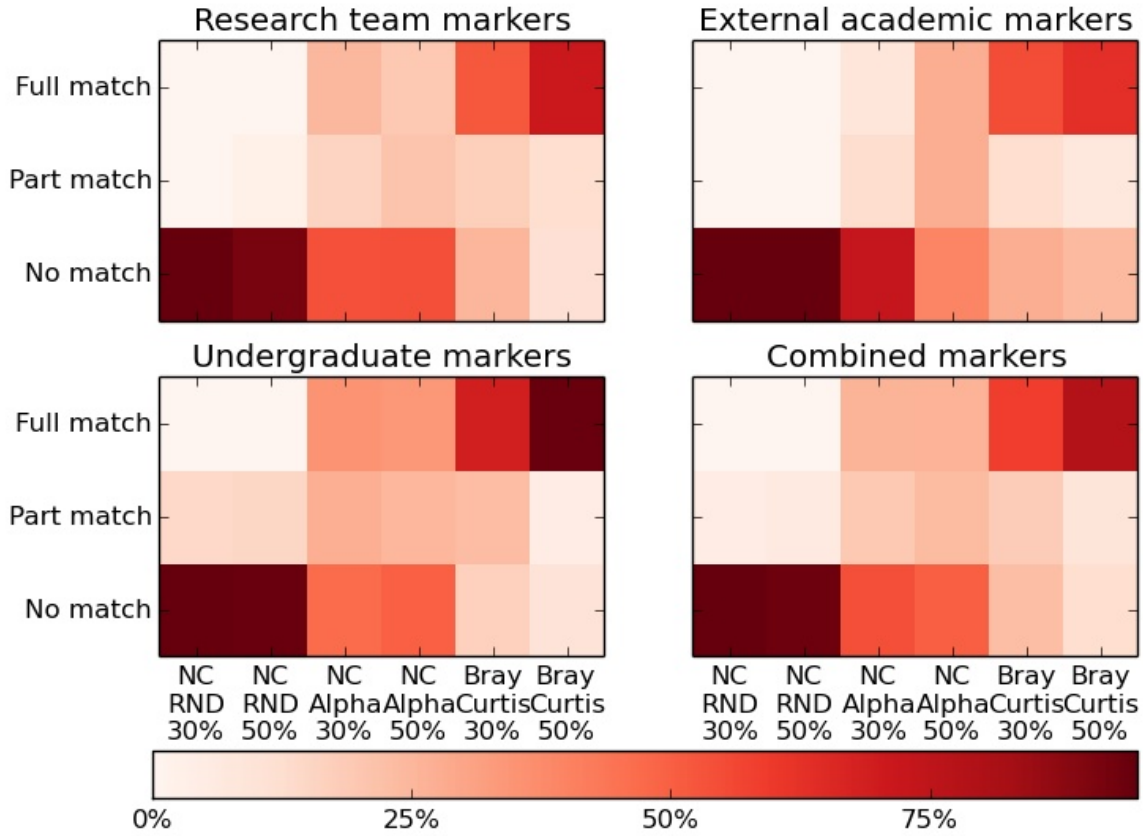


Fig 2. Heatmap showing the similarity in the classification judgements. Distributions of annotations of the outputs (full match—in the output all accounts match all other accounts; part match—some account in the output match other accounts, at least one does not; no match—none of the accounts in the output match any others) from the Bray-Curtis and Random (RND) and Alphabetical (Alpha) control models are visibly similar between all four annotator categories: research team, external academics, undergraduates and combined.

3.4.2 Public profile features

When annotating account, annotators were free to choose their own coding criteria and note down the reasons for matching or not matching accounts. We were able, therefore, to also compare the reasons given by annotators of different expertise levels (Fig 3). The

six most common categories of reason for making an annotation decision were similar across research team, external academic and undergraduate annotators. The reasons given were different, however, when humans annotated the entire dataset of 2,144 accounts. There was no evidence to reject the null hypothesis that reasons given were independent of expertise level (chi-squared test, df =18, $\chi^2 = 0.1200$, p = 1.0).

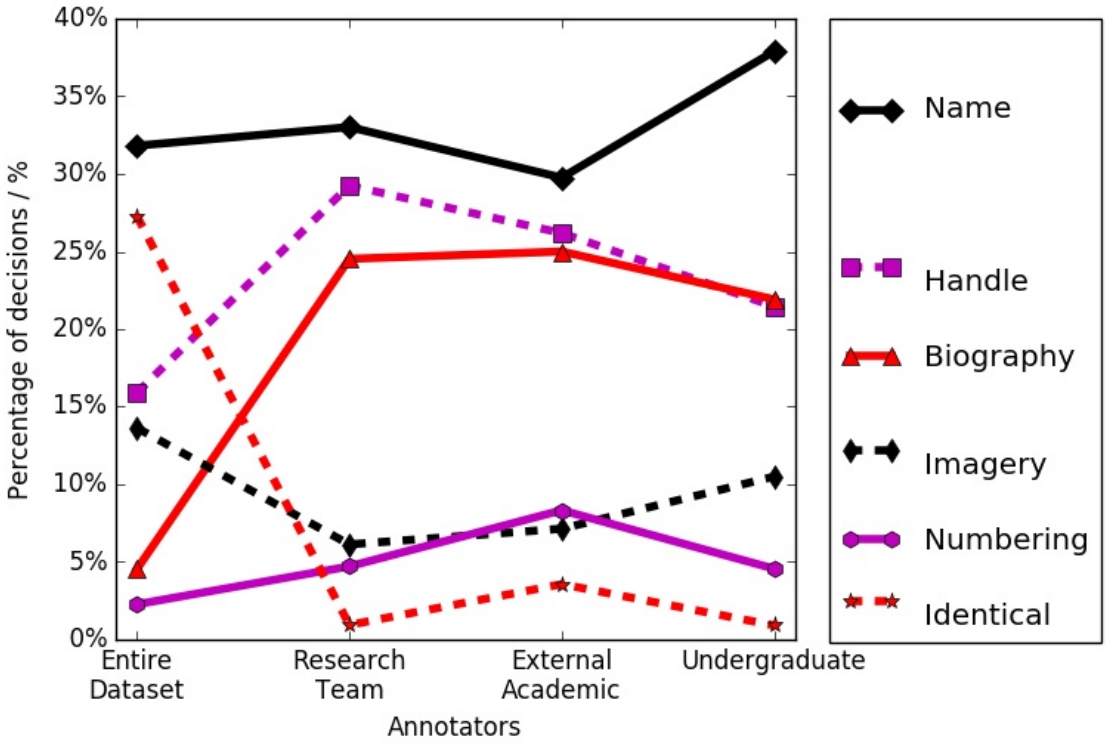


Fig 3. The six most common categories of reason for making an annotation decision were similar across research team, external academic and undergraduate annotators. The reasons given were different, however, when humans annotated the entire dataset of 2,144 accounts.

3.4.3 Images

Having taking a screenshot of each Twitter account in our sample (except for 602 that had been suspended and 97 that no longer existed due to name changes), we investigated further the importance of imagery to humans when annotating the datasets. Two example snapshots from screenshots are given in Fig 4.

Whilst 28.5% of accounts across the entire dataset had a screenshot, pseudoreplicate sets found by human annotation had significantly more (mean = 66%; one-sample, 2-tailed t-test t = 4.63, p = 0.0000623) (Fig 5). Furthermore, whilst there is no evidence (all p > 0.187) that screenshot prevalence in the random and representative negative control model outputs differs from population, both Bray-Curtis models output accounts with significantly fewer screenshots (Student's one-sample, 2-tailed t-test: BC_{30} $p < 6.59x10^{-9}$; BC_{50} $p < 1.97x10^{-12}$) (Fig 5).

Fig 4. A section of the screenshots of two illustrative accounts whose screenshots could be informative—the whole profile picture was presented. Screenshots of two user accounts taken from http://twitter.com.

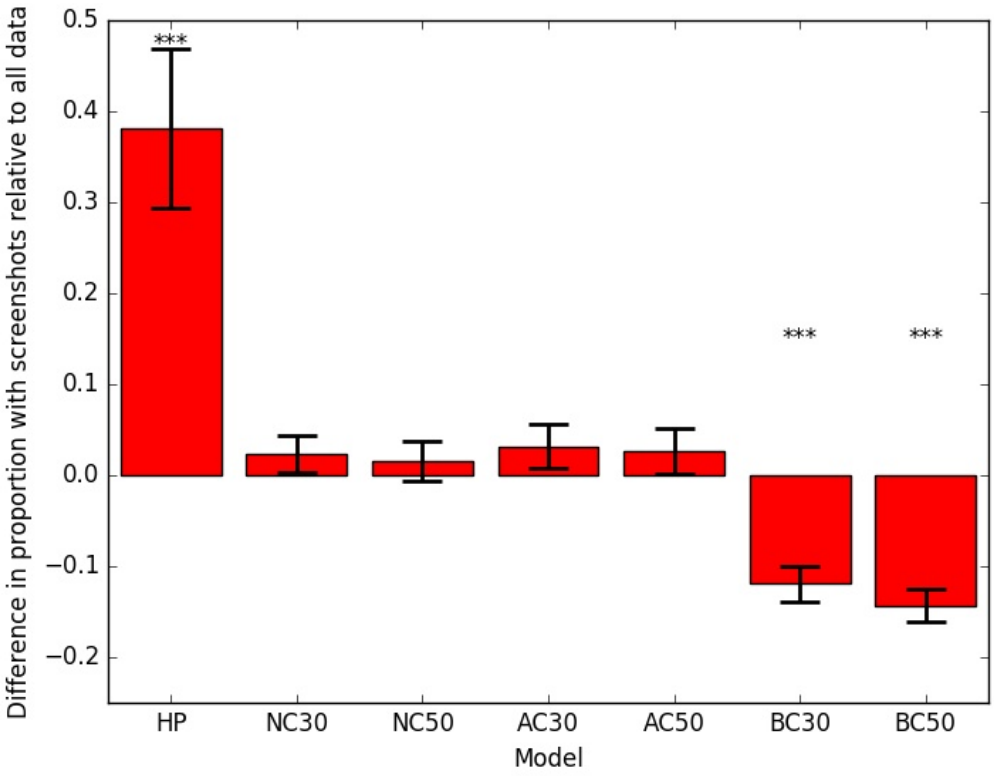


Fig 5. Human annotation relies on screenshots, whilst pseudoreplicate accounts found with machine methods are less likely to have a screenshot. The difference between the mean proportion of accounts with screenshots in sets of pseudoreplicates found by each model and the proportion of accounts with screenshots in the entire dataset (0.285). Error bars ±1 S.E.M. Stars signify significant one-sample t-tests against population; all $p < 0.0001$.

4 Discussion

With social media datasets often constituting tens to hundreds of thousands of accounts, finding and identifying pseudoreplicate accounts—whether to analyse terrorism, marketing trends or other law enforcement—requires machine methods. Here we develop and evaluate a new machine method based on publicly available profile features.

Human-annotation is currently used for finding and matching pseudoreplicate Twitter accounts sampled from the domain of terrorism and, at first, seems to be the logical test-dataset. Here, we have shown, however, that when human-annotation is evaluated against positive controls—accounts that we have downloaded twice and know with certainty to be pseudoreplicates—human annotation fails to find many of them (recall = 9.91%). Human annotation is therefore not a good method for generating test-datasets.

On the other hand, had we not been subject to the ethical limitations of sharing terrorism related content with online crowdsourcing platforms, such as Amazon Mechanical Turk, these findings may not apply. Given that other researchers investigating topics of particular ethical sensitivity might face similar obstacles, this research hopefully demonstrates alternative methods to generate test-data and evaluate methods.

Through *post hoc* annotation of individual outputs, we have also shown that our text similarity based machine method can find and match pseudoreplicates more reliably

than the human annotation (recall = 95.5%; G mean = 0.977; F score = 0.636). Although the model does not perform this well when it is evaluated against human-annotation, we argue that this is because humans annotating the entire dataset suffer more memory and time handicaps than humans annotating individual pseudoreplicates.

As the positive controls were discovered by accident during analysis, after human-annotation of the dataset, we benefited from the unplanned ability to evaluate human-annotation against 'positive controls'. They were not included in the experimental design, however, and this meant that we were unable to compare enough of them. Future work could specifically include a large enough number to calculate how many human-annotation correctly identifies and how many it misses. Further, the metrics that were used could not be directly applied to the machine methods, as accounts were loaded in by Twitter ID and thus only ever loaded in once—i.e. there were no positive controls—whereas a specifically designed study could incorporate them into the design.

In the latter part of our paper, we showed that the classifications given by annotators from different backgrounds were significantly correlated. Further, there was no evidence that reasons given to justify classifications were independent of expertise (chi-squared test, df =18, $\chi^2 = 0.1200$, $p = 1.0$). The most common reasons (name, handle, biography) also used the same features as our machine model, suggesting that the classification of pseudoreplicates is not dependent on subjective opinion, but that the salient information is just as amenable to machine methods.

With individual, *post hoc* annotation of model outputs, however, the reasons given for decisions did not correlate significantly with the reasons given during human-annotation of the entire dataset of 2,144 account. For example, whereas 27% of whole dataset classification-reasons were because accounts were absolutely identical, this was only 2% when individually annotating model outputs. Furthermore, human annotation of the entire dataset was biased. The accounts it found were significantly more likely to have screenshots than accounts in the population as a whole were ($p = 0.0000623$). Not only is human annotation of the entire dataset biased relative to the population on average, but as resurgents get suspended frequently, they are even less likely to exist long enough for screenshot capture than the population is on average. That pseudoreplicate accounts identified by our machine methods were significantly less likely to have screenshots than the population as a whole ($p < 6.59x10\text{-}9$) adds further evidence to support the validity of our machine methods. It also suggests, potentially because memory and time limitations mean humans rely on obvious visual imagery wherever possible (Wright *et al.*, 2016), that human analysis is biased towards finding a particular, unrepresentative subset of resurgents—those with matching profile and background images (although this suggests that image recognition software could potentially be used in future to automate human performance). On the other hand, our model now appears to be a better performing method.

There are other reasons why online accounts need matching up and it would also be interesting to know whether this technique could assist with those too and to what extent these approaches work when some attempt is made by the user at anonymity. One example is the detection of the often forbidden fake profiles (Gurajala, 2015). In this

paper we connect account across longitudinal samples within a network and this could work similarly to other research attempting to connect accounts across social networks (Goga *et al.*, 2015; Korula and Lattanzi, 2014; Malhotra, 2013; Vesdapunt and Garcia-Molina, 2014). There are a variety of reasons this might be useful: law enforcement matching personal information to identify the owner of an account; marketing of products to a user who expresses interest elsewhere; or sociological analysis, for example matching political views expressed in one place to a person's demographic details given elsewhere. There are obvious ethical implications to some of these goals, but our algorithm raises tantalising questions about whether they are possible, as well as highlighting the risks of distributing details about oneself across multiple accounts.

As the only previous approach to find and match pseudoreplicates amongst the tens of thousands of Baqiya family accounts active every month was human annotation—which we have demonstrated has flaws—had led to only five case studies and left analysts of terrorism and intelligence struggling to find and control for resurgents, it is likely that their datasets suffered from pseudoreplication. In this paper, we showed that novel methods, based upon the text similarity of publicly available profile features can quickly and reliably classify Twitter pseudoreplicates within the domain of terrorism.

Acknowledgements

All authors were involved in the conception of the work. SPW collected the data, performed the analyses, wrote the first draft and led the writing of the manuscript. AP, DD, VAAJ and JB edited and critiqued the manuscript. The authors would also like to express their gratitude to Peter Adey for helpful discussions and feedback on the manuscript. SPW is funded by the Royal Holloway University of London Reid Scholarship—security and sustainability theme.

References

1. Amarasingam A. What Twitter Really Means For Islamic State Supporters. War on the Rocks; 2015.
2. Amazon Mechanical Turk [software]. Available at: https://www.mturk.com/mturk/welcome
3. Berger JM, Morgan, J. The ISIS Twitter Census: Defining and describing the population of ISIS supporters on Twitter. The Brookings Institution; 2015
4. Berger JM, Perez H. The Islamic State's Diminishing Returns on Twitter: How suspensions are limiting the social networks of English-speaking ISIS supporters [Occasional Paper]. George Washington Program on Extremism; 2016.
5. Bray JR, Curtis JT. An ordination of the upland forest communities of southern Wisconsin. Ecological monographs. 1957;27(4): 325–349.
6. Bryden J, Funk S, Geard N, Bullock S, Jansen VAA. Stability in flux: community structure in dynamic networks. Journal of the Royal Society, Interface. 2011;8(60): 1031-40.
7. Bryden J, Funk S, Jansen VAA. Word usage mirrors community structure in the online social network Twitter. EPJ Data Science. 2013;2(3)
8. Brynielsson J, Horndahl A, Johansson F, et al. Analysis of Weak Signals for Detecting Lone Wolf Terrorists. 2012 European Intelligence and Security Informatics Conference. 2012; 197-204. DOI 10.1109/EISIC.2012.20
9. Chatfield AT, Reddick CG, Brajawidagda U. Tweeting propaganda, radicalization and recruitment: Islamic state supporters multi-sided twitter

networks. Proc 16th Annual International Conference on Digital Government Research. 2015: 239-49 DOI:10.1145/2757401.2757408

10. Cormack G, Lynam T. TREC 2005 Spam Track Overview. Sixteenth Text Retrieval Conference (TREC 2007). 2005: 1-9
11. Goga O, Loiseau P, Sommer R, Teixeira R, Gummadi KP. On the Reliability of Profile Matching Across Large Online Social Networks. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2015: 1799-1808
12. Goodman LA. Snowball Sampling. Ann Math Stat. 1961;32: 148–170
13. Greenberg J. Why Facebook and Twitter Can't Just Wipe Out ISIS Online. WIRED: Business [newspaper on the Internet]. 21 Nov 2015. Available: http://www.wired.com/2015/11/facebook-and-twitter-face-tough-choices-as-isis-exploits-social-media/. Accessed 22 Apr 2016
14. Gurajala, 2015. Proceedings of the 2015 International Conference on Social Media & Society - SMSociety '15.
15. Harris C, Srinivasan P. Hybrid Crowd-Machine Methods as Alternatives to Pooling and Expert Judgments. Information Retrieval Technology, Airs 2014, LNCS 8870. 2014: 60-72
16. Hurlbert SH. Pseudoreplication and the Design of Ecological Field Experiments. Ecological Monographs. 1984;54(2): 187-211. DOI: 10.2307/1942661
17. Korula, N., Lattanzi, S., 2014. An efficient reconciliation algorithm for social networks. Proceedings of the VLDB Endowment. 7(5):377-388
18. Lo S, Chiong R, Cornforth D. Using support vector machine ensembles for target audience classification on Twitter. PLoS ONE. 2015;10(4): 1-20. DOI: 10.1371/journal.pone.0122855

747 19. Magdy W, Darwish K, Weber I. #FailedRevolutions: Using Twitter to Study the
748 Antecedents of ISIS Support. arXiv preprint. 2015; arXiv:1503.02401v1
749 20. Malhotra A. Studying User Footprints in Different Online Social Networks.
750 arXiv preprint. 2013; arXiv:1301.6870v1
751 21. McPherson M, Smith-Lovin L, Cook JM. Birds of a Feather: Homophily in
752 Social Networks. Annual Review of Sociology. 2001;27: 415-44.
753 22. Powers DMW. Evaluation: from precision, recall and f-measure to ROC,
754 informedness, markedness & correlation. Journal of Machine Learning
755 Technologies. 2011;2(1): 37-63.
756 23. Preoțiuc-Pietro D, Volkeva V. Lampos V, Bachrach Y, Aletras N. Studying User
757 Income through Language, Behaviour and Affect in Social Media. PLoS ONE.
758 2015;10(9): e0138717. doi:10.1371/ journal.pone.0138717
759 24. Rykiel E. Testing ecological models: The meaning of validation. Ecological
760 Modelling. 1996;90(3): 229-244
761 25. Smucker MD, Kazai G, Lease M. Overview of the TREC 2012 Crowdsourcing
762 Track. Proceedings of the 21st NIST text retrieval conference (TREC-2012).
763 2012
764 26. Sokolova M, Japkowicz N, Szpakowicz S. Beyond accuracy, F-Score and ROC:
765 A family of discriminant measures for performance evaluation. Advances in
766 Artificial Intelligence, AI 2006, LNAI 4304. 2006;4304: 1015-1021.
767 27. Stern J, Berger JM. ISIS: The state of terror. London: William Collins; 2015.
768 28. Tamburrini N, Cinnirella M, Jansen VAA, Bryden J. Twitter users change word
769 usage according to conversation-partner social identity. Social Networks.
770 2015;40: 84-89.

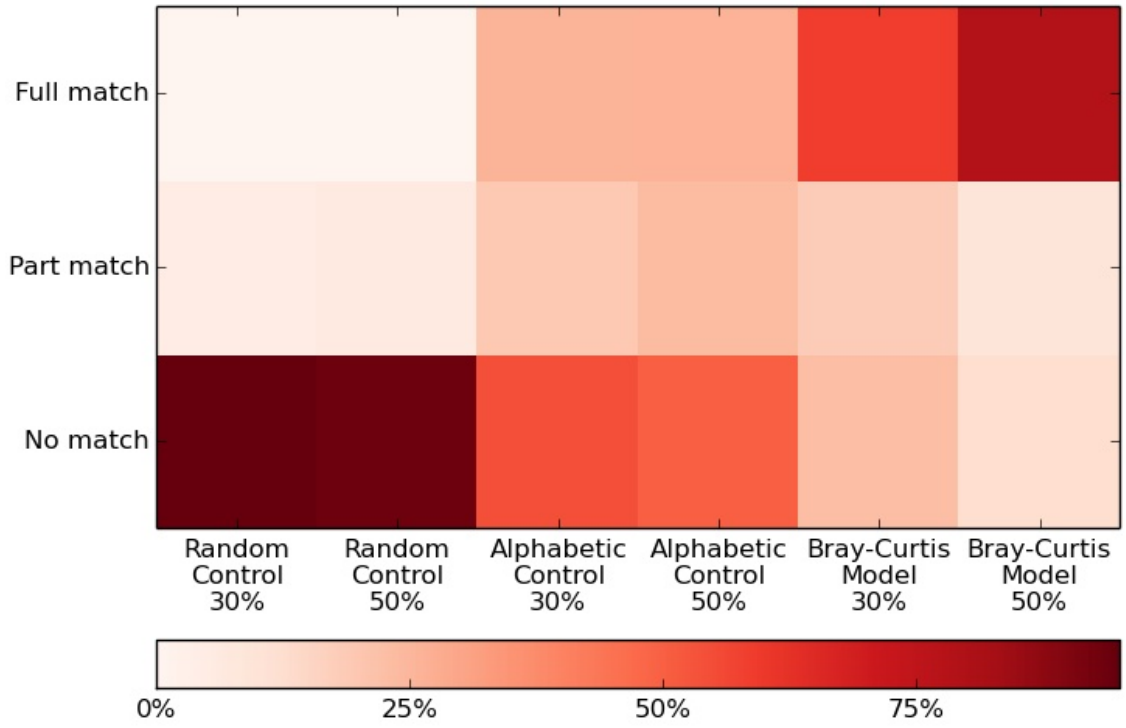
29. Vaux DL, Fidler F, Cumming G. Replicates and repeats—what is the difference and is it significant? A brief discussion of statistics and experimental design. EMBO Rep. 2012;13(4): 291–296. doi: 10.1038/embor.2012.36
30. Vesdapunt N, Garcia-Molina H. Identifying Users in Social Networks with Limited Information. 2014: 1-44.
31. Weimann G. New Terrorism and New Media. Washington, DC: Commons Lab of the Woodrow Wilson International Center for Scholars; 2014.
32. Wright S, Denney D, Pinkerton A, Jansen VAA, Bryden J. Resurgent Insurgents: Quantitative Research Into Jihadists Who Get Suspended but Return on Twitter. Journal of Terrorism Research. 2016;7(2): 1–13. doi: 10.15664/jtr.1213
33. Yang C, Srinivasan P. Translating surveys to surveillance on social media. Proceedings of the 2014 ACM conference on Web science - WebSci '14. 2014: 4-12. DOI: 10.1145/2615569.2615696

Supporting Information

S1 Table.

Test dataset	(a) Human Performance			(b) +Improved Human Annotation		
model	RC_{30}	AC_{30}	BC_{30}	RC_{30}	AC_{30}	BC_{30}
TP	0	3	11	0	91+1+2	476+3
TN	2,043,600	2,043,614	2,043,836	2,046,268	2,045,790	2,044,378
FP	3636	3622	3400	968	1355	2390
FN	40	37	29	40	37	29
TPR / sensitivity / recall	0.00%	7.50%	27.5%	0.00%	71.8%	94.3%
TNR / specificity	99.8%	99.8%	99.8%	100%	99.9%	99.9%
Accuracy	99.8%	99.8%	99.8%	100%	99.9%	99.9%
Precision	0.00%	0.0828%	0.3225%	0.00%	6.49%	16.7%
G mean	0.00	0.274	0.524	0.00	0.847	0.970
F score	0.00	0.00164	0.00637	0.00	0.119	0.284
LAM	#DIV/0!	0.304	0.235	#DIV/0!	0.143	0.112

Performance of three models: Random Control (30%), Alphabetic Control (30%) and Bray-Curtis (30%). All results given to 3.s.f.



790

791 **S1 Fig.** Heatmap showing the similarity of the performance of the 30% and 50%
792 models, as evaluated by improved human annotation. 95% and 94% of the randomly
793 matched, control sets were judged as no match, whilst 59% and 78% of sets matched on
794 the basis of Bray-Curtis similarity were judged as full match.

795

7. Bickering Families: an Analysis of the Baqiya Family on Twitter and Offline Daesh Events

Overview in relation to the thesis

The fourth and final research chapter tackles three related questions. Firstly, is it possible to infer information about events in the offline world by tuning in to online data? Secondly, is this possible in data from the field of terrorism studies? While attempts have been made to show that it is (Magdy *et al.*, 2015), sufficient controls and excluding non-terrorism data have not been used. Finally, a much broader question is addressed—are various computational and discourse metrics still statistically reliable when applied to noisy, big data from social media.

To investigate these issues, three computational and two discourse analysis metrics are evaluated as predictors, from Daesh Twitter data, of the parity of geopolitical events relating to Daesh. This work demonstrates that whilst a small number of results remain statistically significant, the majority of results that emerge from the data can just as easily arise by chance in the negative control data. This raises potential questions about the validity of approaches such as sentiment analysis, computational metrics and discourse analysis, or alternatively, highlights issues with defining negative control data amongst subjective, ambiguous or patchy data.

Wright, S., Denney, D., Pinkerton, A., Jansen, V.A.A., 2016. Bickering Families: an Analysis of the Baqiya Family on Twitter and Offline Daesh Events. [*In preparation*]

Bickering Families: an Analysis of the Baqiya Family on Twitter and Offline Daesh Events

Shaun Wright, Alasdair Pinkerton, David Denney, Vincent AA Jansen

Abstract

Tweeting patterns about Daesh are thought to differ depending on the nature of offline events involving Daesh. This paper replicates previous findings more rigourously and novelly shows that Daesh tweets more on positive day—rather than tweeting a larger proportion of positive tweets. We also demonstrate, however, that computational methods are susceptible to identifying apparently significant patterns in noisy, negative control data, and thus particular care should be taken when drawing conclusions from computational or human coded methods.

Introduction

Terrorism supporters on Twitter—their "favourite" social network (Weimann, 2014)—are predominantly (in 2015) supporters of Daesh, with tens of thousands of accounts (Berger and Morgan, 2015). Daesh supporters form a largely 'grass-roots' community known as 'the Baqiya family' (Amarasingam, 2015; Huey and Witmer, 2016; Miller, 2015): "a loose network of Islamic State supporters from around the world who share news, develop close friendships, and help each other" (Amarasingam, 2015). This large volume of Twitter data has enabled the application of a range of computational (Berger and Morgan, 2015; 2016; Magdy *et al.*, 2015; 2016; Ferrara, 2016; Rowe, 2016) and discourse-based (Klausen, 2015; Stern and Berger, 2015; Ghajar-Khosravi *et al.*, 2016; Huey and Witmer, 2016) based approaches to terrorism—historically, a data-sparse phenomenon (Silke, 2009).

One such novel, computational approach was the monitoring of attitudes towards Daesh in relation to real-world events (Magdy *et al.*, 2015). Magdy *et al.* showed that days where Twitter had net positive sentiment towards Daesh, offline news related to Daesh contained "the release of propaganda videos and major military achievements" (Magdy *et al.*, 2015). On the other hand, on days where net sentiment was negative, "news of

[Daesh] human rights violations emerged, such as the killing of hostage [sic], accounts of torture, or reports of the enslavement of Yazidi women" (Magdy *et al.*, 2015). This appears to suggest a relationship between online sentiment towards Daesh and offline events.

Part One – Confirming that online sentiment towards Daesh matches offline events relating to Daesh

In Magdy *et al.*'s study (2015), however, online and offline data was correlated *post hoc*, subject to confirmation bias and lacking in negative controls. The primary aim of this study is to more rigourously test the hypothesis that online sentiment towards Daesh matches offline events relating to Daesh. To do this, following the standard machine learning or model validation approach, we compare independently gathered datasets relating to our online model and offline events—rather than just, as Magdy *et al.* (2015) appeared to, confirming that 'a' news article exists for each of the model outputs.

In particular, Magdy *et al.* (2015) defined days depending on whether Daesh achieved victories (positive days for Daesh) or suffers losses (negative days for Daesh). We will, therefore, do the same. As big data is noisy and ambiguous, however, we will extend the categories to include control days—neutral (neither victories nor losses for Daesh) and mixed (at least one victory and one loss for Daesh) days. Crucially, we will carry out all of this analysis independently of analysing Twitter sentiment.

The other reason we argue that a re-testing of Magdy *et al.*'s hypothesis is important, is that their results did not demonstrate a relationship between offline events and Twitter sentiment towards Daesh within the Baqiya family—i.e. Daesh's 'internal' popularity. Since their findings came from a cross-section of the whole of Twitter (Magdy *et al.*, 2015; 2016), net negative sentiment appeared to be driven by a significant increase in tweets from users unconnected to Daesh drowning out the pro-Daesh tweets of Daesh supporters (Magdy *et al.*, 2015; Ferrara *et al.*, 2016). Given that the majority of the world condemns Daesh, that result is not surprising. By studying the net sentiment, whether the sentiment amongst Daesh supporters changes is not reported. For several reasons, however, including ensuring the accuracy of terrorism analysis and providing feedback on counter-narratives, we consider it important to investigate whether the

Daesh supporting community also responds differently to positive and negative offline events. Thus, in this study, we investigate whether Daesh's real-world victories and losses are or are not related to the sentiment of tweets by Daesh supporters—i.e. the Baqiya family.

Finally, by using an English speaking dataset, we test whether the above approach extends into languages other than Arabic (Magdy *et al.*, 2015; 2016).

Part Two – Subgroup of the Baqiya family

Assuming such a relationship between online sentiment and offline events can be demonstrated, methods could be developed to allow academics and law enforcement agencies to track events before they get reported in the media. Further, such methods could also provide a feedback loop to governments that are attempting to degrade Daesh's capabilities and popularity.

Historically—when al-Qaeda was dominant—the jihadist landscape was highly fractionated into intermittently bickering, defecting, funding and allying affiliates. One of the key issues for security services was understanding the relationships between these subgroups and predicting when they were about to team up or divorce from one another, although understanding the important connections between them was difficult even for the intelligence agencies (Storm *et al.*, 2003). The techniques discussed in the first section—using online sentiment to assess offline events and behaviours—appear to present a possible way to understand the strength of such inter-group relationships.

Under the unifying caliphate of the so-called Islamic State, however, the Baqiya family of 2015/16 has few, if any, factions or franchises (Amarasingam, 2015). While this would appear to limit the applicability of the methods in this paper, we would argue that understanding inter-group relationships is a significantly important problem (ultimately and indirectly, lives could be saved), that it is worth testing for any subgroup stratification that might informatively exist.

To that end, while there remain some al-Qaeda and Syrian opposition factions, their presence does not form a large proportion of jihadist Twitter, nor are they highly interlinked with the Baqiya family. One logical partitioning of the Baqiya family we

would argue, therefore, is geographical heritage. Thus, in the second part of this paper, we develop the analysis further and investigate whether subgroups based on geographical and cultural background influences the way Baqiya family subgroups respond to Daesh's offline-world activities. In particular, we look at whether each geographical subgroup follows the same tweeting patterns, or whether subgroups respond to events differently.

Part Three – Metric comparisons

In the preceding sections of this paper, we assess sentiment on Twitter using Magdy *et al.*'s heuristic (Magdy *et al.*, 2015; 2016). Their heuristic is based on their finding that "the full name of group is a strong indicator of support for ISIS (93%), and using the acronyms is a general indication of opposition (77%)" (Magdy *et al.*, 2015; 2016). In the final part of our study, we investigate other sentiment assessment metrics reveal the same patterns.

As a range of competing approaches emerge, with corresponding benefits and limitations, from different methodological fields, we develop and compare computational methods against those based on content analysis. Computational approaches, first, can handle larger datasets and thus provide higher statistical power. Programmed to test *a priori* hypotheses, however, often means discarding additional, contextual information. On the other hand, methods based on human coding of content and discourse—incorporating not only what is said, but by who, in what context and with what meaning—can detect more nuanced details that may not have been predicted. Such analysis can therefore, provide a more detailed, higher resolution analysis. Content analysis is, however, if done manually, time-expensive and rate-limited to much smaller datasets. Here, we test two new computational heuristics (one calculating the similarity of the words spoken, and the other the similarity of the overall sentiment of the text). We also test two human coded analysis of sentiment methods (one assessing sentiment towards Daesh and the other towards a range of geopolitical organisations). Given the time expensiveness of manually coded analysis, however—and ethical barriers to using online crowdsourcing platforms—we focus on a shorter time period surrounding a specific event: the Sousse 2015 terrorist attacks in order to annotate a complete dataset.

In this study we test whether—like Twitter as a whole—the Baqiya family expresses different sentiment towards Daesh on days that are independently coded as positive (major achievements), negative (major losses), neutral (neither) or mixed (both) for Daesh. To investigate whether inter-group relationships are still amenable despite the homogeneity of the Baqiya family, we then test how different geographical subgroups amongst the Baqiya family respond. Finally, we compare the results from four additional computational or human coding-based metrics and provide detailed analysis of the weeks surrounding the Sousse 2015 attacks.

Methods

Dataset

We sampled 1,920 jihadist, jihadist supporting, or jihadist linked Twitter accounts; the same dataset sampled and characterised in Wright *et al.* (2016). We snowball sampled (Goodman, 1961) between May and July 2015 using the Twitter API. We seeded sampling with 34 English-speaking, jihadist accounts identified through manual analysis of Twitter, aided by Twitter's "Who to follow" suggestions. To build a highly intra-linked Twitter community, we weighted the snowball sampling. This is based on the principle of homophily: the tendency of people to associate with similar people (McPherson *et al.*, 2001) and bias their interactions to members of the same community with whom they share a social identity (Bryden *et al.*, 2011; 2013; Tamburrini *et al.*, 2015). We therefore added, daily, any account followed by >10% of the users already in our sample, and with <1,000 followers of its own. We also downloaded the entire daily tweet output of our sample, generating a corpus of approximately 155,000 tweets. This methodology and rationale is described further in Wright *et al.* (2016).

Classifying day types

Magdy *et al.* (2015) classified positive days for Daesh where Daesh experienced major victories. As we are attempting to replicate their conclusions, we adopt a similar experimental design, but extend it to include control days: neutral (neither victories nor losses/defeats) and mixed (at least one victory and one loss).

We identified real world events of interest during our sampling period (15th May 2015 to 13th July 2015) using the well-sourced "Timeline of ISIL-related events (2015)" (Timeline of ISIL-related events (2015), Wikipedia). We classified all events from the timeline as 'positive' or 'negative' for Daesh (Table 1) and then classified each day as 'positive', 'negative', 'mixed' or 'neutral' depending on the types of events that occurred on that day. 'Positive' days had only positive events, 'negative' days had only negative events, 'mixed' days had at least one positive and one negative event, whilst 'neutral' days had neither type of event reported (Table 2). The implications of this design are discussed further in the limitations section.

Table 1. Numbers of different events relating to Daesh.

Type of event (for Daesh)	Category of event	Count
Positive	Daesh captures/retakes entity position	9
	Daesh or affiliate terror attack / suicide bomb /kidnapping	9
	Daesh attacks/clashes with entity	2
	Daesh punishes / executes / destroys	5
	Evidence against al-Baghdadi death rumour	1
Negative	Entity captures/retakes Daesh position	4
	Member(s) of Daesh killed	15
	Daesh failed clashes/attack against entity	4
	Entity attacks/clashes with Daesh	2
	Airstrikes against Daesh	3
	US arrest for providing support to Daesh	3
	US deploys anti-Daesh advisors to Iraq	1

The events (and numbers of them) that occurred between 15th May - 13th July 2015 categorised and classified as positive or negative for Daesh.

Table 2. Numbers of days that were positive, negative, mixed or neutral for Daesh.

Type of day (for Daesh)	Definition	Count (% days)
Positive	At least one positive event. No negative events.	15 (25%)
Negative	At least one negative event. No positive events.	19 (32%)
Mixed	At least one positive, and at least one negative, event.	6 (10%)
Neutral	No positive events and no negative events.	20 (33%)

The days (and numbers of them) that occurred between 15th May - 13th July 2015 categorised and classified as positive, negative, mixed or neutral for Daesh.

Part One – Confirming that online sentiment towards Daesh matches offline events relating to Daesh

We tested whether the sentiment expressed by the whole Baqiya family was significantly different on the four day types (positive, negative, mixed or neutral) (see Methods—classifying day types).

We used the method outlined in the introduction from Magdy *et al.* (2015; 2016), but as our sample was English, rather than Arabic, speaking, we adapted new markers of pro or anti-Daesh sentiment (Table 3).

Table 3. Strings indicating sentiment towards Daesh.

Sentiment (towards Daesh)	Case	Word strings
Positive	insensitive	"I.S."; "Baqiya"; "Baqiyah"; "Caliphate"; "Islamic State"
	sensitive	"IS"
Negative	insensitive	"ISIS"; "ISIL"; "Daesh"

Word strings classed as markers of positive or negative sentiment towards, Daesh.

After determining day types as outlined above, for each of the positive days, we used the Magdy heuristic to calculate the absolute number of positive references to Daesh, the absolute number of negative references, and the proportion of positive / negative references as scaled by the overall number of tweets made on that day. We then repeated the process for the other day types. As the data was non-normal (Kolmogorov-Smirnov test), we used the Kruskal-Wallis non-parametric, one-way ANOVA (4.s.f.) to test for differences in the distributions of sentiment across the four days. We also used non-parametric Mann-Whitney U tests (4.s.f.), controlling for multiple hypothesis testing with a Bonferonni correction factor of 6, to test for significant differences between pairs of day types.

Controls

As a control, we randomly allocated the 60 days to four negative control groups (to match the four day types). We then repeated the above statistical analysis to establish whether any significance could be achieved by chance in the negative control data. As a

bootstrap we repeated the process (n=20,000) and recorded the proportion of Kruskal-Wallis, and Mann Whitney U, tests that were significant at the $p<0.05$, and $p<0.01$, level.

Part Two – Subgroups of the Baqiya family

We also tested whether the sentiment expressed was significantly different on the four day types for each Baqiya family subgroup—identified as follows:

To categorise the country-of-origin of each profile, we manually inspected the content of their name, biography and location. As multiple, overlapping national identities are common, we followed a coding scheme (Table 4), prioritising in a hierarchy which types of information were most important and overruled others. For example, for an account with "al-Britani" in the name and currently living in "Syria", the coding scheme prioritises the country-of-origin in the name (Table 4, rank 1). On the other hand, for a self-described "Somali-American", the main nationality (American) was prioritised over the heritage (Table 4, rank 2). Only when no geographic information was given were generic categories such as "Internet" (Table 4, rank 4) or "Baqiya family" (Table 4, rank 5) used.

Table 4. Coding scheme—hierarchy of affiliations given by accounts

Rank	Affiliation
1	Country-of-origin given in the name or username, e.g. "al-Britani", "al-Kanadi", "al-Amriki", etc. (*most significant*).
2	Self-described nationality, e.g. "British" (*nationality overriding heritage, e.g. "Somali American"*).
3	Self-declared location, e.g. "Location: Turkey", "@ShamFighter".
4	Sarcastic response or joke about their location, e.g. "Behind you", "Under your bed", "Blockistan".
5	Affiliation generically with the 'Baqiya family'.
6	Affiliation generically with Daesh, e.g. "IS", "Islamic State", "Caliphate", "Dawla Islamiya", "ISIS", "ISIL".
7	No location given.

The hierarchical priority of information about affiliation given by sampled accounts.

As a result of the non-uniform group sizes generated by the first round of country/organisation classification, small subgroups (approximately <20) from similar geographical, cultural and political (all three) backgrounds were repeatedly merged in such a way as to create groups approximately equal in both size and culture.

For each subgroup, we downloaded 60 corpora of tweets—one for each of the 60 days of sampling. For the 27 subgroups, this gave a total of 1,620 unique corpora representing the tweets of a single group on a single day. The corpora are unique and non-overlapping in that no tweet will be present in more than one corpora (unless uniquely re-tweeted on multiple days or by multiple subgroups).

To calculate the sentiment of a subgroup towards Daesh on a given day, we again used the Magdy heuristic. As well as the absolute number of positive references and the number of negative references to Daesh, we scaled them by dividing scaled by the total number of tweets in the corpora for that day.

We repeated the whole process for each of the 27 subgroups. We used a Bonferonni correction factor of 27 to control for the multiple Kruskal-Wallis tests; making $p_{significance} < 0.00185$.

Part Three – Metric comparisons

For two subgroups, we calculated the similarity of their corpora on each day—using the Magdy heuristic to calculate the proportion of positive references to Daesh in each and then calculating a similarity by subtracting the absolute difference between these (mathematically equivalent to using negative references) from one. We then calculated the mean similarity for each day type (positive, negative, neutral and mixed). We repeated this for each of the 351 pairwise combinations of the 27 Baqiya family subgroups. Using Student's t-tests (Bonferroni corrected, 2-tailed, subgroup-paired, 4.s.f.), we compared mean similarities on different day types: positive days with negative; positive with neutral; positive with mixed; negative with neutral; negative with mixed; and neutral with mixed. As in part one, we also randomly allocated days to control groups (n=4) and compared with Bonferroni corrected t-tests.

We then repeated the analysis with two new computational metrics: Bray-Curtis similarity and a Sentiment analysis calculator.

Bray-Curtis similarity

The Bray-Curtis index (Bray and Curtis, 1957) is a standard metric in ecology and in computer science for comparing the similarity of two populations of animals or words. It calculates (twice) the proportion of words (or species) shared by both populations. After counting the number of shared words in the lowercase version of the corpora, this is divided by the total number of words (Formula 1). Importantly, multiple instance of the same word are treated as independent instances.

Formula 1.

$$Bray-Curtis\ similarity = \frac{2 \cdot number\ of\ word\ shared\ by\ both\ populations}{Total\ number\ of\ words\ across\ both\ populations}$$

Sentiment analysis calculator

Open source algorithms are widely available to assess the sentiment of a block of text. They usually return a continuous score between +1 (completely positive) and -1 (completely negative). Our basic model analyses the text as a whole, although more complex analyses are possible where sentiment with respect to a given keyword is assessed. We used AlchemyAPI's free Sentiment Analysis API (AlchemyAPI, IBM Watson). We compared the daily corpora of two subgroups by submitting the two to AlchemyAPI's API and calculating the absolute difference between the returned sentiment scores. We then scaled this between 0 and 1.

Sousse attack

To introduce human coding of sentiment models, we looked at the weeks before and after the 26th June 2015, Sousse terrorist attack. On the 26th June 2015 Seifeddine Rezgui attacked tourists at a beach front hotel in Sousse, Tunisia. Before the arrival of security forces, the gunman, with links to Daesh, killed 38 people—including 30 British tourists—and wounded 39 others. Four other Islamist attacks took place on the same day in France, Kuwait, Syria and Somalia. To do this, we repeated Part Two (a), using only the days between the 19th June and 2nd July 2015. This gave us time to annotate and include the two human coded analysis of sentiment metrics when comparing pairs of subgroup's daily corpora.

Human coding of sentiment towards Daesh

For each subgroup, we downloaded 14 corpora of tweets—one for each of the seven days before and seven after Sousse. For the 27 subgroups, this gave a total of 378 non-overlapping corpora. We presented each corpora in a separate word document, with the text of the tweets in chronological order. To avoid confirmation bias, however, we anonymised each document blind to both group and day. The number of tweets in each daily corpus were similar across most groups (median=27, mean=40), with the exception of the daily corpora of the "No location" group (median=800, mean=788). We randomly selected 100 tweets from these corpora for analysis. The authors read each tweet and, using open criteria, coded whether it referred to Daesh. We noted the sentiment of any reference to Daesh also (positive, negative, neutral). We then calculated the total number of positive, negative and neutral tweets referencing Daesh in each corpus. For statistical analysis, we used the proportion of a subgroup's daily tweets that were pro-Daesh.

Human coding of overall sentiment

While coding references to Daesh, we also coded references to any other group/individual/entity mentioned. An organically growing list of 39 entities thus emerged (Supplementary Table 1). For each entity, we calculated the proportion of each subgroups' daily tweets that referenced it positively. To calculate the similarity of two daily subgroups' corpora, we calculated the 39 dimension Euclidean distance (e_d), although the count for sentiment towards other groups were often low. To match the other metrics, we then scaled this by the maximum possible difference to create a similarity, rather than distance, measure.

We first tested whether overall similarity changed. For each of the 351 pairs of subgroups, we calculated the mean similarity in the seven days before Sousse and the mean in the seven days afterwards, then tested for differences between the combined before and after data (2-tailed, paired, Student's t-test). We repeated this with each of the five metrics.

We also tested for differences as a results of Sousse in each of the 351 pairs of subgroups individually (2-tailed, unpaired, Student's t-test), again with each of the five metrics.

Controls

When we repeated the analysis for individual pairs of subgroups, we corrected for the 351 multiple hypothesis tests with a Bonferonni correction factor of 351; making $p_{significance} < 0.000142$.

To provide additional baseline similarities with the computational models, we also used five further weeks preceding the Sousse attack, comparing consecutive weeks with 2-tailed, paired Student's t-tests.

Similarly to earlier in the paper, we generated a bootstrapped, negative control by shuffling the 14 days around Sousse and randomly assigning them to two control conditions. We repeated this (n=100), calculating the mean p-value of re-samples (2-tailed, unpaired, Student's t-test).

Results

We sampled 1,920 jihadist, jihadist supporting, or jihadist linked Twitter accounts; the same dataset sampled and characterised in Wright *et al.* (2016). We also downloaded the entire daily tweet output of our sample, generating a corpus of approximately 155,000 tweets. By the end of sampling, 1,080 had been suspended, 141 accounts were private, 97 no longer existed and 602 were still active. The majority of accounts do not declare a terrorist organisation affiliation, although of the 13% that do, all gave ISIS, IS, Islamic Caliphate, Baqiya or Khilifa. Twitter also suspended 56.3% of our sample, evidence that suggests they were engaging in extremist activity. We therefore categorise our sample as jihadist-linked, while assuming, based on location and content, that the majority are Daesh-supporting members of the Baqiya family (Amarasingam, 2015).

Part One – Confirming that online sentiment towards Daesh matches offline events relating to Daesh

Of the 60 days between 15th May and 13th July 2015, 15 were positive for Daesh, 19 were negative, 6 were a mixture of positive and negative, and 20 were neutral, with neither positive nor negative events for Daesh (Figure 1; Supplementary Figure 1).

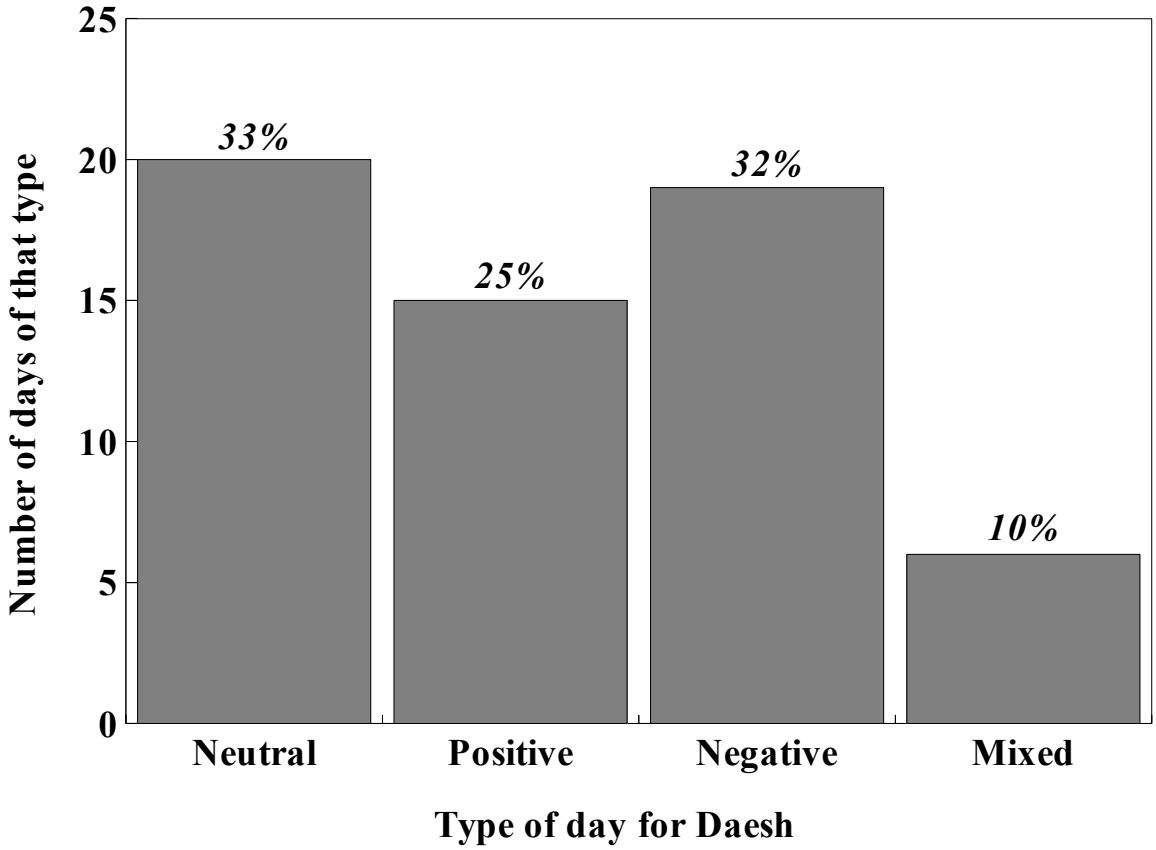


Figure 1. The number of days that Daesh experienced as positive, negative, neutral and mixed between 15th May and 13th July 2015.

Using the Magdy heuristic, we tested whether Baqiya family sentiment towards Daesh was significantly different across positive, negative, mixed and neutral day types.

Absolute number of positive references

The number of positive references to Daesh ("I.S."; "Baqiya"; "Baqiyah"; "Caliphate"; "Islamic State", "IS") was significantly different across day types (Kruskal-Wallis, 4.s.f., $p = 0.0001526$) (Figure 2; Figure 3). Bootstrapped, negative controls were only significant at the expected rate of random chance (0.06055 at $p<0.05$; 0.00095 at $p<0.001$).

All individual comparisons between pairs of day types were significant—except neutral/positive—even after controlling for the twice-chance rate derived from the bootstrapped, negative control (0.1138 at $p<0.05$; 0.0238 at $p<0.01$): positive/negative (Mann-Whitney U test, 4.s.f., $p = 0.002944$), positive/mixed (Mann-Whitney U test,

4.s.f., p = 0.00003992), negative/mixed days (Mann-Whitney U test, 4.s.f., p = 0.02078), negative/neutral (Mann-Whitney U test, 4.s.f., p = 0.02446), neutral/mixed (Mann-Whitney U test, 4.s.f., p = 0.0001812) (Figure 2; Figure 3).

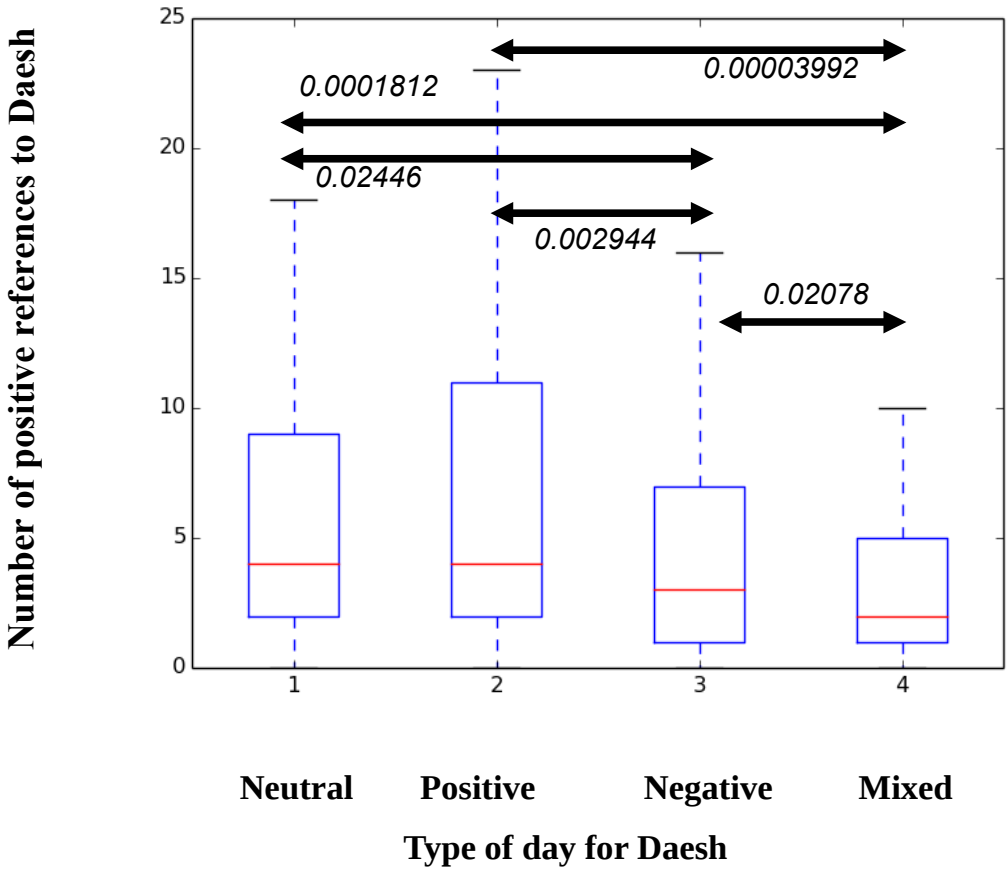


Figure 2. Box and whisker plots showing the positive references to Daesh on different types of day (neutral, positive, negative and mixed) is significantly different. P-values of significant (Bonferroni corrected) Mann Whitney U tests are overlaid between days.

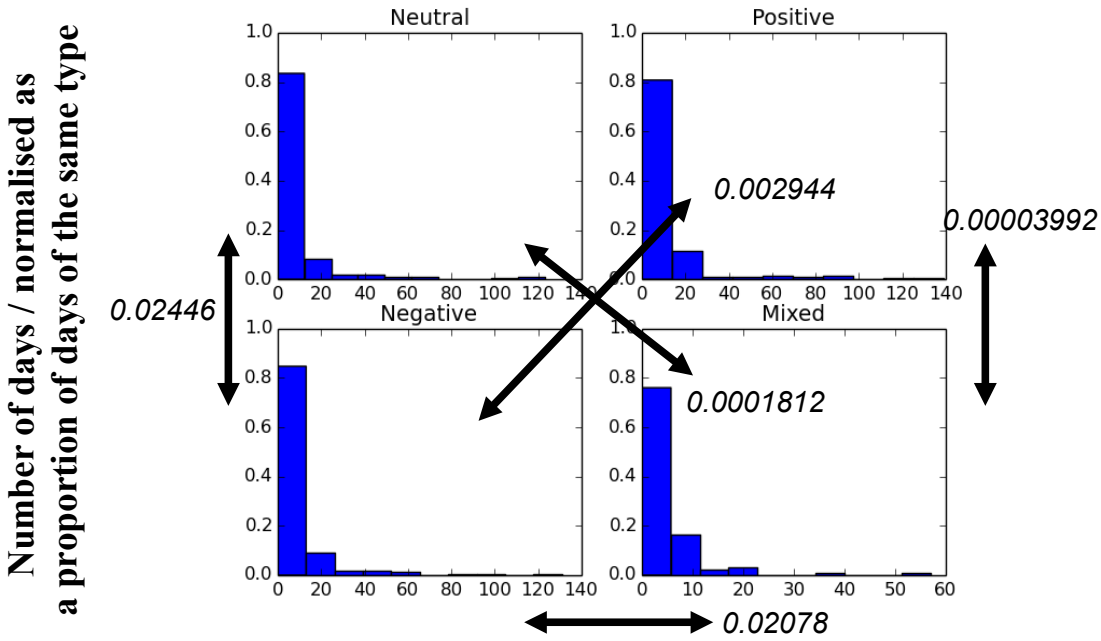


Figure 3. Normalised histograms showing positive references to Daesh on different types of day (neutral, positive, negative and mixed) is significantly different. P-values of significant (Bonferroni corrected) Mann Whitney U tests are overlaid between days.

Proportion of positive references

We scaled by the number of tweets made the same day to test whether the increase in positive references to Daesh is being driven by more mentions, or a greater proportion of positive mentions. There was no significant difference between days (Kruskal-Wallis, 4.s.f., p = 0.1434) (Supplementary Figure 2), suggesting that the Baqiya family simply tweets more about Daesh depending on day type.

Absolute number of negative references

The number of negative references to Daesh ("ISIS"; "ISIL"; "Daesh") was not significantly different across day types (Kruskal-Wallis, 4.s.f., p = 0.1112) (Supplementary Figure 3).

Part Two – Subgroup of the Baqiya family

Categorising the accounts using the coding scheme gave 27 subgroups (Table 5); also represented graphically in Figure 4.

Table 5. Geographical subgroups in our sample.

Subgroup	Number of accounts	Coding scheme rank(s)
No location	837 (100 randomly selected)	7
Daesh	155	6
Baqiya family	100	5
Maghreb	87	1, 2, 3
“Dar ul Kufr” (land of the unbelievers)	83	3
UK	74	1, 2, 3
Sham (Iraq/Syria)	68	1, 2, 3
Sarcasm	65	4
Somalia	63	1, 2, 3
Entire Biography Blank	58	7
Western Europe	53	1, 2, 3
“The World”	41	3
“The Internet”	34	3
Arabian Peninsula	33	1, 2, 3
Africa (other than Maghreb)	32	1, 2, 3
North America (US / Canada)	31	1, 2, 3
South East Asia	30	1, 2, 3
Dunya (This world, as opposed to the next)	24	3
Khorasan	22	3
Turkey	19	1, 2, 3
Eastern Europe / Russia	18	1, 2, 3
Kashmir / India	17	1, 2, 3
Kurdistan	16	1, 2, 3
“Dar al Islam” (land of Islam)	16	3
Lebanon	14	1, 2, 3
Australia	14	1, 2, 3

Afghanistan	13	1, 2, 3

The distribution of our sample by 27 geographical origin subgroups, based on self-identified information. The priority with which they were categorised (Table 4) is given also.

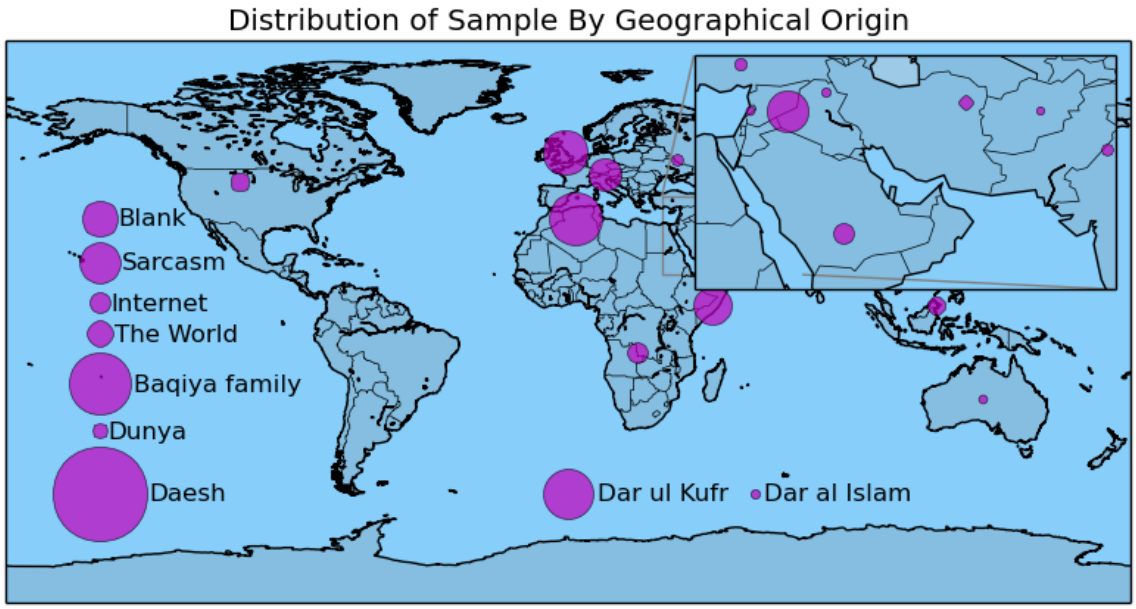


Figure 4. Graphical representation of the distribution of our sample by geographical origin. The area represents the number of Twitter accounts self-identifying as from that category of location; for scale the 'Baqiya family' circle represents exactly 100 users. Although the inset map of the Middle East is magnified by a factor of x2.5, the circles remain unaffected and are thus directly comparable. Users giving no information (n=837) are excluded from the map. Markers within the boundaries of a single country represent that country, markers overlapping several countries represent a group of several neighbouring countries.

None of these subgroups differed significantly across day types in the number, or proportion, of either positive or negative references to Daesh (Bonferroni corrected, Kruskal-Wallis, all $p > 0.05$).

Part Three – Metric comparison

We tested whether the mean similarity of pairs of subgroups of the Baqiya family differed between pairs of day types with three computational metrics. Although the Magdy heuristic suggested groups, on average, differed significantly on positive days compared to mixed days (2-tailed, paired, t-test, p = 0.0008913), one of the negative controls also reached significance (2-tailed, paired, t-test, p = 0.03968). Thus, we found no statistically significantly evidence—relative to negative controls—for differences in sentiment between subgroups of the Baqiya family between any types of day (Figure 5). Nor did we find any evidence that any individual pair of Baqiya family subgroups differed significantly between any day types.

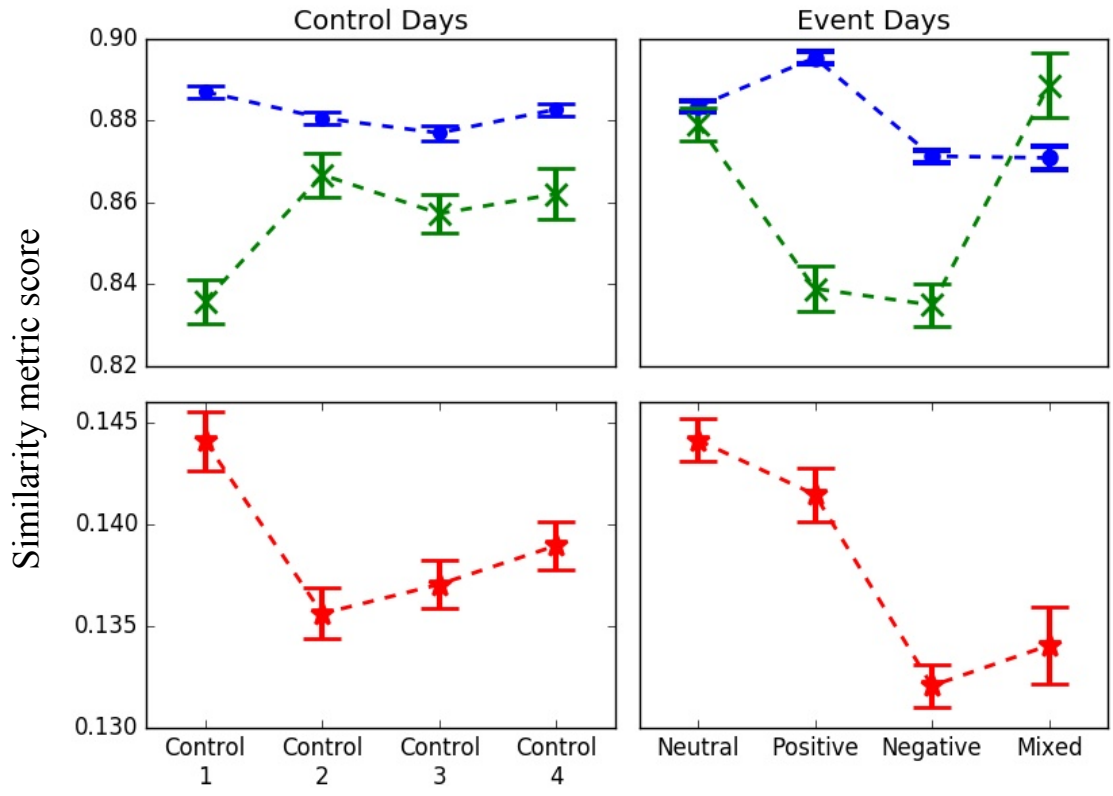


Figure 5. Mean difference between subgroups of the Baqiya family on four types of event day (right hand panels) and four randomly shuffled controls (left hand panels) as measured with three different computational metrics: Bray-Curtis similarity (red stars, bottom panels), difference in online sentiment analysis calculator scores (blue circles, top panels), difference in proportion of references to Daesh that were pro-Daesh (green crosses, top panels). Error bars represent ±1 S.E.M.

Sousse attack

We looked at the mean difference between subgroups of the Baqiya family to see if, in the week after the Sousse attack (26th June–2nd July 2015), there was a significant change in their mean similarity relative to the week before the attack (19th–25th June 2015). For each pair of groups, we calculated the similarity of their tweets on each of the 14 days represented here.

Model A – Bray-Curtis

Subgroups were significantly less similar in Bray-Curtis sentiment in the week before Sousse (0.130) than in the week afterwards (0.138) (t-test, $p < 0.001$; Cohen's d = 0.1082 (Very Small)) (Figure 6). Groups shared significantly more words in common in the week after the Sousse attack than in the week before it. There was no significant difference between weeks in the bootstrapped, negative control (mean $p = 0.536$). However, as can be seen in the figure, these results are not statistically significant relative to the noise—with larger effect sizes—in the preceding control weeks.

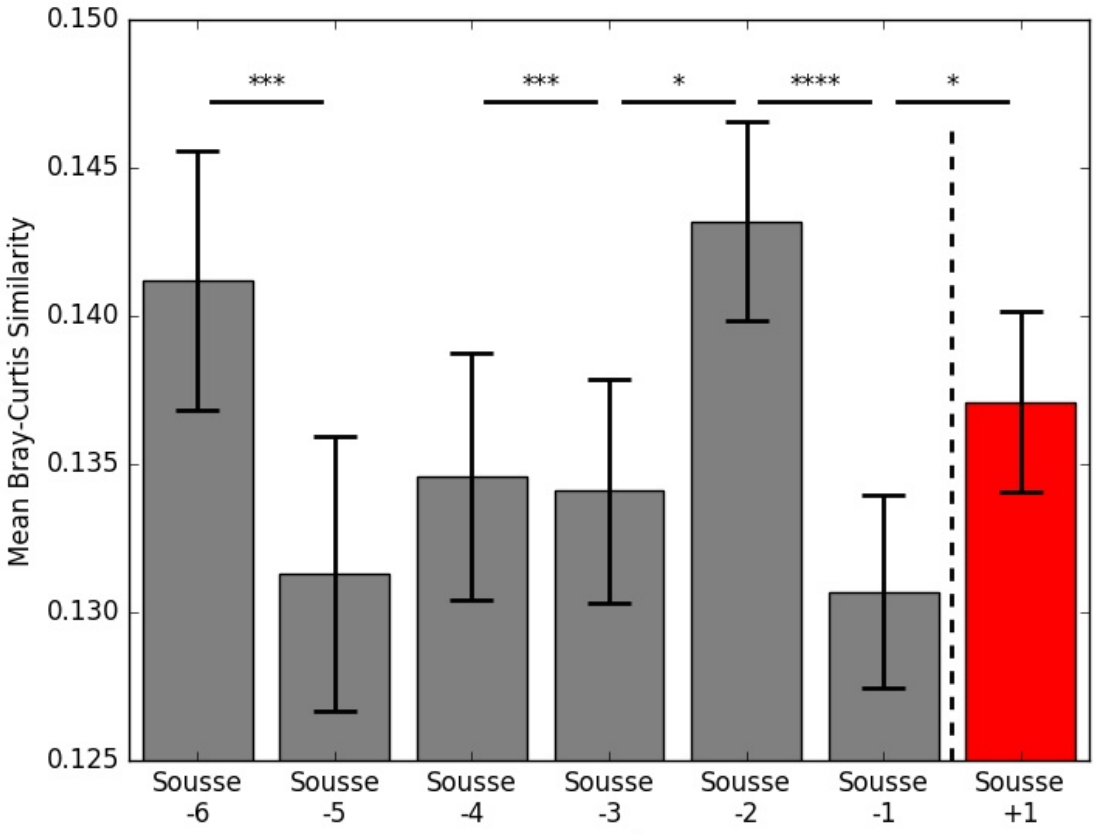


Figure 6. Mean difference between subgroups of the Baqiya family in the weeks before (grey bars) and after (red bar) the 2015 Sousse attacks (dashed line), as measured with the Bray-Curtis metric. Significant differences between consecutive weeks are indicated by stars: * $p \leq 0.05$; ** $p \leq 0.01$; *** $p \leq 0.001$; **** $p \leq 0.0001$ (Bonferonni corrected, 2-tailed, paired Student's t-test). Error bars represent ±1 S.E.M. Cohen's d effect sizes for pairwise comparisons left to right are: 0.122 (Very Small); 0.0416 (Very Small); 0.000688 (Very Small); 0.136 (Very Small); 0.201 (Small); and 0.108 (Very Small).

None of the individual subgroup pairs differed significantly different across the weeks.

Model B – Sentiment Analysis

Subgroups were significantly more similar in 'sentiment analysis' sentiment in the week before Sousse (0.08) than in the week afterwards (0.141) (t-test $p < 1.48x10^{-12}$; Cohen's d = 0.4908 (Medium)) (Figure 7). The sentiment of groups' tweets became significantly more different in the week after the Sousse attack than in the week before it. There was no significant difference between weeks in the bootstrapped, negative control (mean p = 0.514). However, as can be seen in the figure, these results are not significant relative to the noise—albeit noise with smaller effect sizes—in the preceding control weeks.

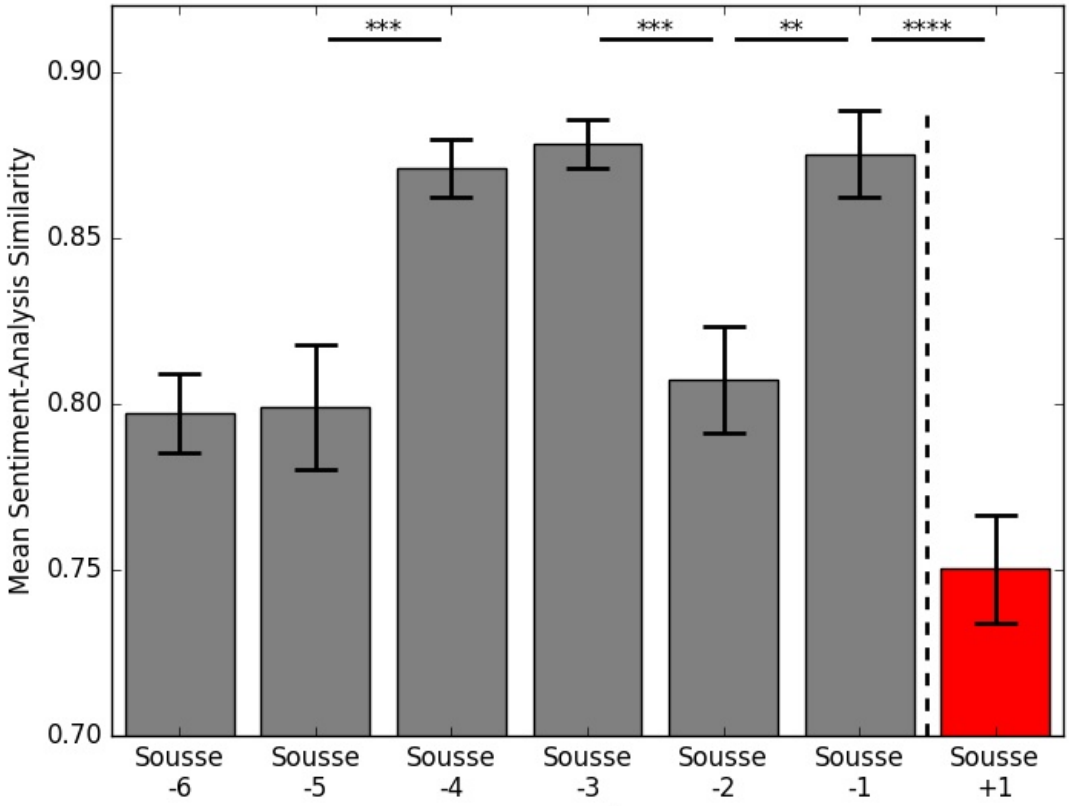


Figure 7. Mean difference between subgroups of the Baqiya family in the weeks before (grey bars) and after (red bar) the 2015 Sousse attacks (dashed line), as measured with the online sentiment calculator metric. Significant differences between consecutive weeks are indicated by stars: * $p \leq 0.05$; ** $p \leq 0.01$; *** $p \leq 0.001$; **** $p \leq 0.0001$ (Bonferonni corrected, 2-tailed, paired Student's t-test). Error bars represent ±1 S.E.M. Cohen's d effect sizes for pairwise comparisons left to right are: 0.00811 (Very Small); 0.358 (Small); 0.0501 (Very Small); 0.352 (Small); 0.276 (Small); and 0.491 (Medium).

None of the individual subgroup pairs differed significantly different across the weeks.

Model C – Magdy Heuristic

Subgroups were significantly less similar in Magdy heuristic sentiment in the week before Sousse (mean = 0.868) than in the week afterwards (mean = 0.885) ($p < 4.54 \times 10^{-7}$; Cohen's d = 0.2982 (Small)) (Figure 8). Significantly fewer of groups' references to Daesh were of the same parity in the week after the Sousse attack than in the week before it. There was no significant difference between weeks in the bootstrapped, negative control (mean p = 0.490). However, as can be seen in the figure, these results are not statistically significant relative to the noise—with larger effect sizes—in the preceding control weeks.

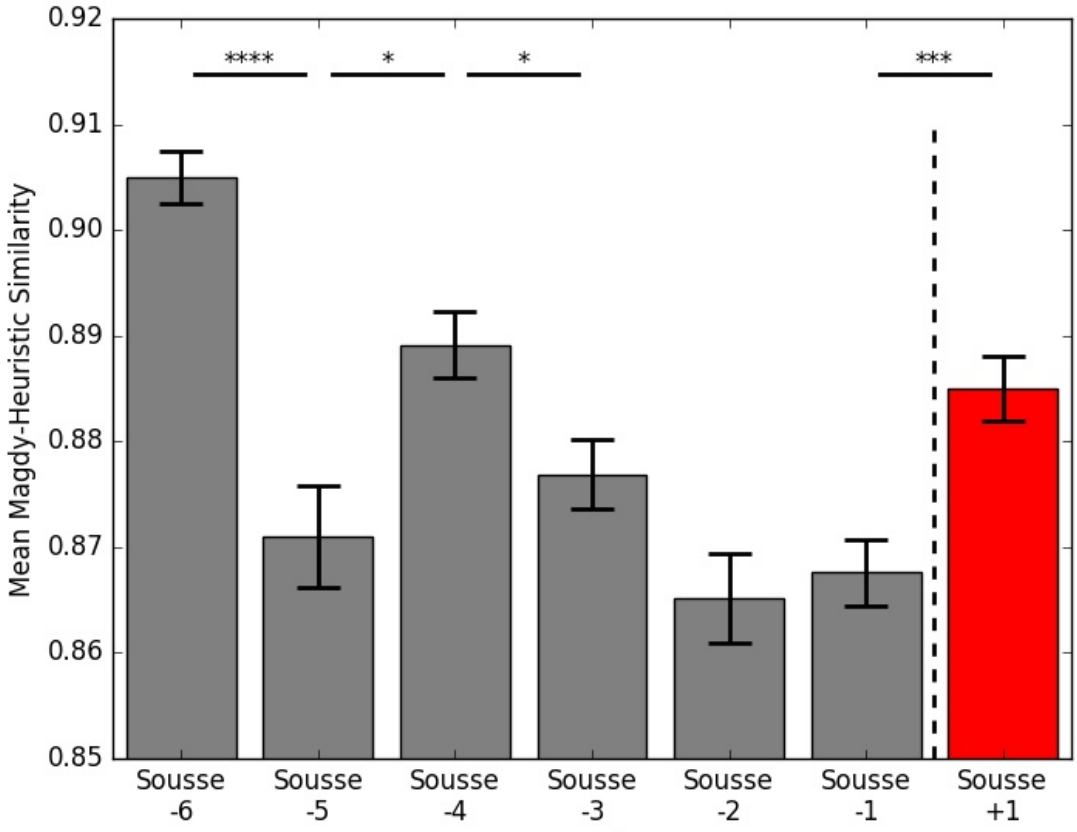


Figure 8. Mean difference between subgroups of the Baqiya family in the weeks before (grey bars) and after (red bar) the 2015 Sousse attacks (dashed line), as measured with the Magdy heuristic metric. Significant differences between consecutive weeks are indicated by stars: * $p \leq 0.05$; ** $p \leq 0.01$; *** $p \leq 0.001$; **** $p \leq 0.0001$ (Bonferonni corrected, 2-tailed, paired Student's t-test). Error bars represent ±1 S.E.M. Cohen's d effect sizes for pairwise comparisons left to right are: 0.493 (Medium); 0.248 (Small); 0.213 (Small); 0.169 (Very Small); 0.0348 (Very Small); and 0.298 (Small).

The following subgroup pairs had a significantly different mean similarity in the week after the Sousse attack relative to their mean similarity in the week preceding it (Table 6). There was no significant difference between weeks in the bootstrapped, negative control (all $p > 0.05$).

Table 6. Groups whose Magdy Heuristic similarity changed significantly over Sousse.

Group One	Group Two	Mean (*s.d.*) group similarity in the week before Sousse	Mean (*s.d*) group similarity in the week after Sousse	Significance (2-tailed, unpaired, equal variance Student's t-test)	Effect size (Cohen's d)
Sham	Africa (other)	0.874 (*0.195*)	0.730 (*0.139*)	$p < 0.000116$	0.863
Sham	Eastern Europe/Russia	0.880 (*0.192*)	0.713 (*0.160*)	$p < 1.62 \times 10^{-5}$	0.952
Sham	Kurdistan	0.883 (*0.191*)	0.712 (*0.157*)	$p < 6.12 \times 10^{-6}$	0.986
Sham	Afghanistan	0.888 (*0.188*)	0.713 (*0.154*)	$p < 1.46 \times 10^{-6}$	1.02
Sham	Kashmir/India	0.881 (*0.192*)	0.698 (*0.165*)	$p < 6.08 \times 10^{-7}$	10.3‡
Sham	Internet	0.877 (*0.198*)	0.701 (*0.165*)	$p < 1.48 \times 10^{-6}$	0.971
Sham	Lebanon	0.868 (*0.206*)	0.700 (*0.163*)	$p < 3.65 \times 10^{-6}$	0.917
Sham	No location	0.865 (*0.206*)	0.709 (*0.159*)	$p < 3.91 \times 10^{-6}$	0.861
Sham	Khorasan	0.861 (*0.209*)	0.699 (*0.168*)	$p < 1.42 \times 10^{-6}$	0.866
Sham	dar al Islam	0.863 (*0.207*)	0.699 (*0.168*)	$p < 8.14 \times 10^{-7}$	0.875
Sham	The World	0.858 (*0.211*)	0.702 (*0.166*)	$p < 1.41 \times 10^{-6}$	0.833
Sham	Baqiya	0.858 (*0.212*)	0.704 (*0.161*)	$p < 5.79 \times 10^{-7}$	0.827
Western Europe	dar al Islam	0.975 (*0.0829*)	0.865 (*0.146*)	$p < 0.000112$	0.946
Western Europe	Baqiya	0.979 (*0.0758*)	0.879 (*0.142*)	$p < 3.48 \times 10^{-5}$	0.888
Internet	Baqiya	0.972 (*0.0522*)	0.681 (*0.163*)	$p < 1.13 \times 10^{-5}$	3.31‡

Pairs of groups—from comparisons of 351 pairwise combinations of groups with Model C—with significantly different similarity after Sousse 2015. Results reported to 3.s.f. All effect sizes are Large, with two Gigantic‡.

Model D – Human coded sentiment analysis (Daesh references)

Subgroups were not significantly different in their human coded sentiment (towards Daesh) in the week before Sousse (similarity = 0.187) than in the week afterwards (similarity = 0.138) (2-tailed, paired t-test, p = 0.312). There is no evidence that subgroups diverged or converged in the proportion of their tweets that were pro- or anti-Daesh.

There was insufficient data for statistical analysis of individual pairs of subgroups.

Model E – Human coded sentiment analysis (Euclidean similarity)

Subgroups were not significantly different in their human coded sentiment (Euclidean) in the week before Sousse (similarity = 0.987) than in the week afterwards (similarity = 0.989) (2-tailed, paired t-test, p = 0.501). There is no evidence that subgroups diverged in the overall proportions of tweets they made about various entities.

There was insufficient data for statistical analysis of individual pairs of subgroups.

Whereas the previous model looks at a subset of the content analysis data (referencing Daesh), we can test the subset of the data referencing each of the 39 entities (Supplementary List 1) to check for significant change from the week before to the week after the Sousse attack. None of the 39 entities had significantly different numbers of tweets about them in the week after Sousse compared to the week before (Bonferonni correction applied, $p_{significance} < 0.00128$).

Discussion

Knowing when and why attitudes towards Daesh are changing is important. This is especially true of attitudes towards Daesh from within the Baqiya family. Knowing when attitudes in the Baqiya family change could reveal the most controversial topics and the weak-link (most susceptible to changing their opinions) members of the Baqiya family, thus suggesting counter-narratives and divisive psychological operations (PsyOps) (Garner, 2010). Monitoring attitudes then also acts as a feedback loop on the counter-narratives' effectiveness. Finally, changing attitudes could reveal when Daesh factions are combining resources to plan attacks, exhausting them in battling one another, or about to fracture further into more violent spin-offs, helping governments with ensuring security and forming foreign policy.

We have replicated findings (Magdy *et al.*, 2015) that patterns of tweets about Daesh differ and relate to the type of offline, Daesh-related events occurring that day. By focusing on a Baqiya family-linked dataset and using an independently determined list of events and days, we have reinforced and extended the work into the Baqiya family's internal sentiment, demonstrating that change in sentiment is not exclusively driven by opponents of Daesh being more vocal on specific types of day. If members of the Baqiya family have a source of information other than the media, then there is potential for analysis of Twitter to provide a faster indication of events than waiting for traditional media; this merits future work.

This result presents far from the complete picture, however. When we scaled positive references to Daesh by the number of tweets, no evidence of differences remained. This presents novel evidence, therefore, that suggests that the driving mechanism by which supporters of Daesh express differing sentiment is via the volume of their tweets about Daesh, rather than the parity (positive/negative sentiment) of them.

In the second part of the study, we applied a range of other metrics. Each of these was based on the hypothesis that there would be changes in sentiment. Thus, it is perhaps unsurprising, given our previous finding that only the volume of tweets changes, that these metrics were unsuccessful in finding any significant differences. More concerning, however, is the inability to properly define control data against which to validate

findings. For example, we assumed that the six weeks preceding the Sousse attacks were control weeks where no attacks took place were valid controls—given the ever changing nature of political events, this assumption may not be true. This is in part due to the ambiguous nature of social and political data where developments are continually happening.

In the final section of this paper, we show that although there are statistically significant differences between the weeks before and after the Sousse attacks according to all three computational models, it is only by comparing sufficiently many other weeks that we can see that these statistically significant results are not necessarily sociologically meaningful. It is possible that real-world Daesh-related events happened in these weeks, and so they were not really control weeks, or it is possible that the metrics are just so sensitive to noisy data that they also picked up significant changes in what were supposed to be the preceding control weeks. Furthermore, although this human coding of sentiment found no significant results, and is still too time consuming to check the baseline noise level, we suggest that these are the very reasons why caution should be taken with the future application of human coding of big data. Subjective interpretation of human coded results is liable to suffer from this limitation to an even greater extent. As such methods are unlikely to analyse more than the experimental data they collect, rather than control data that is unrelated to the social topic of interest (e.g. in our study, the preceding week), attempts should be made to falsify the findings relative to more background noise; attempts to demonstrate the validity and sociological meaningfulness of any findings.

Limitations

As a consequence of the ambiguity and lack of an objectively annotated dataset, several assumptions and generalisation have had to be made. Although careful thought was given to ensure that an appropriate choice was made at each point, there remain clear qualifications to this work.

Firstly, the subgroups were artificially demarcated, and only a primary affiliation was selected, despite obvious instances of overlapping. Secondly, the sample adopted from Wright *et al.* (2016) was sampled from those more than 10% connected to previously identified accounts. It is therefore conditional on the accuracy of the initial seed list.

Thirdly, we use a light, quantitative version of content analysis, anonymised and thus blind to some of the context that more traditional social scientists would require. Whilst its simplicity may not satisfy those academics, and may indeed have prevented it from finding any significant results in this study, we argue that it was the only available approach to make it directly testable against the computational metrics.

Finally, the list of offline events relating to Daesh is a secondary resource, with no published methodology for its compilation. It may, therefore, have missed some events, although we can eliminate false positives by confirming each reference given for the events included. Alternative methods proposed for constructing such a list were not practical given the time constraints of the remainder of the work and, even as a naïve list, generated one significant result in our study. Future work with a more fine tuned list may improve both the quality and number of significant results. Further, we subjectively classified the events on the list depending on whether we thought they were victories or losses for Daesh. This depends heavily on assumptions of what makes a day positive or negative for Daesh. This could potentially be difficult to assess (for example air strikes 'could' be a positive marker that Daesh is posing a significant threat to State powers), however we operated on the assumption that Daesh's aims are to seize land, kill Western/Kurdish fighters and not to lose land or Daesh fighters—with the one exception of through suicide attacks, which are clearly a desirable aim from their point of view). Different subjective classifications could clearly have affected our findings, as would any other completely different system to divide the days—it is, however, significant that, while a control splitting the days randomly did not lead to any significant results, classifying by type of offline events did.

An alternative way to frame this study might have been to ask 'what kind of sentiment is provoked amongst Daesh supporters on Twitter by different offline events'. That would have prevented the need to absolutely classify events as positive or negative for Daesh. This would present a sensible future step, having looked in particular at victories and losses, however, in this study we aimed to replicate, as closely as appropriate, the Magdy *et al.* (2015) study.s

We were also unable to replicate any of the above findings between any individual pairs of subgroups of the Baqiya family in any section of the analysis. As previously

discussed, this could be because our subgroups were too artificial, but it could also be that they were too small, or that there was insufficient data overall. This finding does, however, confirm existing work that has shown greater homogeneity in the Baqiya family than under previous terrorism-landscapes (e.g. the period when al-Qaeda was dominant). Future work could exclude any users who are part of multiple subgroups in order to prevent their homogeneity biasing the inter-group comparisons. Future work could also develop an alternative study design that compares Daesh supporters to a control, non-Daesh group. Although that would not address the inter-group splits and team-ups between terrorist groups that are crucial for government and security agencies to understand, it might be more successful (this work has demonstrated an inability to achieve the more desirable aim) and would present a more rigourous—and potentially more sensitive—method for tracking sentiment in the Baqiya family (as sentiment will be appropriately controlled against a baseline).

Conclusion

The Baqiya family do respond differently on Twitter on different types of day, but this is by tweeting more about Daesh, rather than increasing the proportion of tweets that are positive about Daesh.

We cannot show whether computational metrics are able to reliably detect changes in sentiment as we cannot reliably evaluate their results compared to the noise in control periods and human coded methods are too time consuming to generate enough data to detect the subtle changes.

References

Amarasingam, A., 2015. What Twitter Really Means For Islamic State Supporters. *War on the Rocks.* December 30.

Berger, J.M., Morgan, J., 2015. The ISIS Twitter Census: Defining and describing the population of ISIS supporters on Twitter. *The Brookings Institution.*

Berger, J.M., Perez, H., 2016. The Islamic State's Diminishing Returns on Twitter: How suspensions are limiting the social networks of English-speaking ISIS supporters [Occasional Paper]. *George Washington Program on Extremism.*

Bray, J.R., Curtis, J.T., 1957. An ordination of the upland forest communities of southern Wisconsin. *Ecological monographs.* **27**(4):325–349.

Bryden, J., Funk, S., Geard, N., Bullock, S., Jansen, V.A.A., 2011. Stability in flux: community structure in dynamic networks. *Journal of the Royal Society, Interface.* **8**(60):1031-40.

Bryden, J., Funk, S., Jansen, V.A.A., 2013. Word usage mirrors community structure in the online social network Twitter. *EPJ Data Science.* **2**(3)

Ferrara, E., Wang, W., Varol, O., *et al.*, 2016. Predicting online extremism, content adopters, and interaction reciprocity. *arXiv preprint*. ArXiv:1605.00659

Garner, G., 2010. Case Studies in Exploiting Terrorist Group Divisions with Disinformation and Divisive/Black Propaganda. *Journal of Terrorism Research*. **1**(1). DOI: http://dx.doi.org/10.15664/jtr.164

Ghajar-Khosravi, S., Kwantes, P., Derbentseva, N., Huey, L., 2016. Quantifying Salient Concepts Discussed in Social Media Content: A Case Study using Twitter Content Written by Radicalized Youth. *Journal of Terrorism Research*. **7**(2):79–90. DOI: http://dx.doi.org/10.15664/jtr.1241

Goodman, L.A., 1961. Snowball Sampling. *Ann Math Stat.* **32**:148–170

Huey, L., Witmer, E., 2016. #IS_Fangirl: Exploring a New Role for Women in Terrorism. *Journal of Terrorism Research*. **7**(1):1-10. DOI: http://dx.doi.org/10.15664/jtr.1211

IBM Watson. *AlchemyAPI [computer software].* [Available at: <http://www.alchemyapi.com/>] [Accessed: January 2016]

Klausen, J., 2015. Tweeting the Jihad: Social Media Networks of Western Foreign Fighters in Syria and Iraq. *Studies in Conflict & Terrorism*. **38**:1–22. DOI: 10.1080/1057610X.2014.974948

Magdy, W., Darwish, K., Weber, I., 2015. #FailedRevolutions: Using Twitter to Study the Antecedents of ISIS Support. *arXiv preprint.* arXiv:1503.02401v1

Magdy, W., Darwish, K., Weber, I., 2016. #FailedRevolutions: Using Twitter to study the antecedents of ISIS support. First Monday. **21**(2):1-15

McPherson, M., Smith-Lovin, L., Cook, J.M., 2001. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology.* **27**:415-44.

Miller, N., 2015. Global terror 'family' Baqiya a growing concern for security. *The Sydney Morning Herald*. 3rd October. [Available at: <http://www.smh.com.au/world/global-terror-family-baqiya-a-growing-concern-for-security-20151002-gk0fq5.html>] [Accessed: 13th January 2016]

Rowe, M., Saif, H., 2016. Mining Pro-ISIS Radicalisation Signals from Social Media Users. *Association for the Advancement of Artificial Intelligence.*

Silke, A., (ed.), 2009. *Terrorists, Victims and Society: Psychological Perspectives on Terrorism and its Consequences.* Chichester: John Wiley & Sons Ltd.

Stern, J., Berger, J.M., 2015. *ISIS: The state of terror.* London: William Collins.

Storm, M., Cruickshank, P., Lister, T., 2003. *Agent Storm: A Spy Inside al-Qaeda* (3rd ed.). London: Penguin.

Tamburrini, N., Cinnirella, M., Jansen, V.A.A., Bryden, J., 2015. Twitter users change word usage according to conversation-partner social identity. *Social Networks*. **40**:84-89.

Timeline of ISIL-related events (2015), (n.d.). In *Wikipedia.* Retrieved May 25, 2016, from https://en.wikipedia.org/wiki/Timeline_of_ISIL-related_events_(2015)

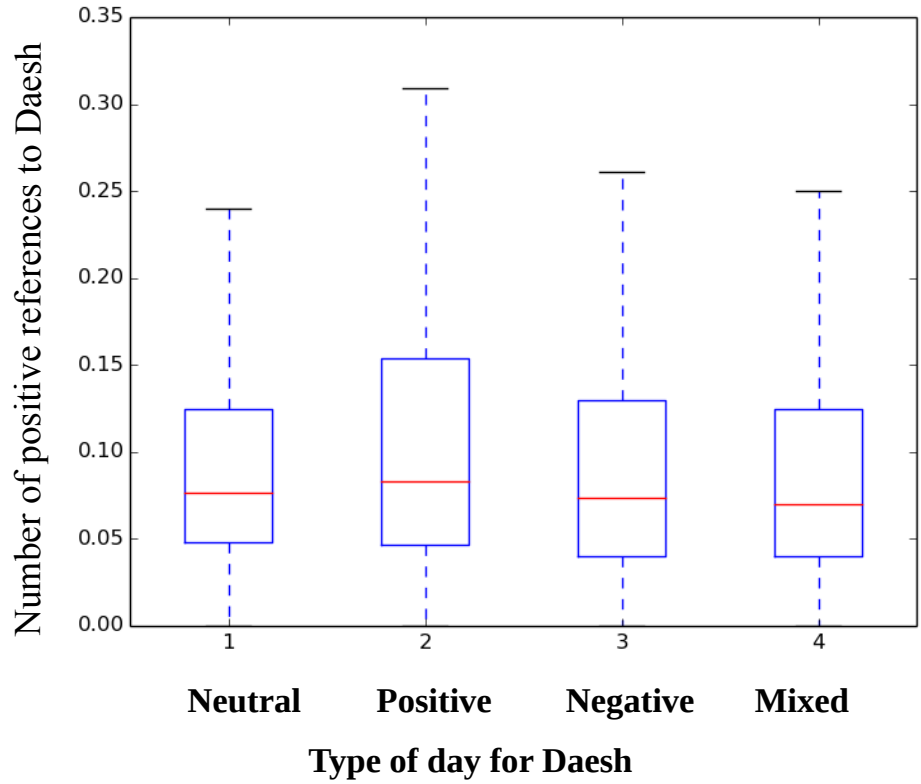
Weimann, G., 2014. *New Terrorism and New Media*. Washington, DC: Commons Lab of the Woodrow Wilson International Center for Scholars.

Wright, S., Denney, D., Pinkerton, A., Jansen, V.A.A., Bryden, J., 2016. Resurgent Insurgents: Quantitative Research Into Jihadists Who Get Suspended but Return on Twitter. *Journal of Terrorism Research.* **7**(2):1–13. DOI: http://dx.doi.org/10.15664/jtr.1213

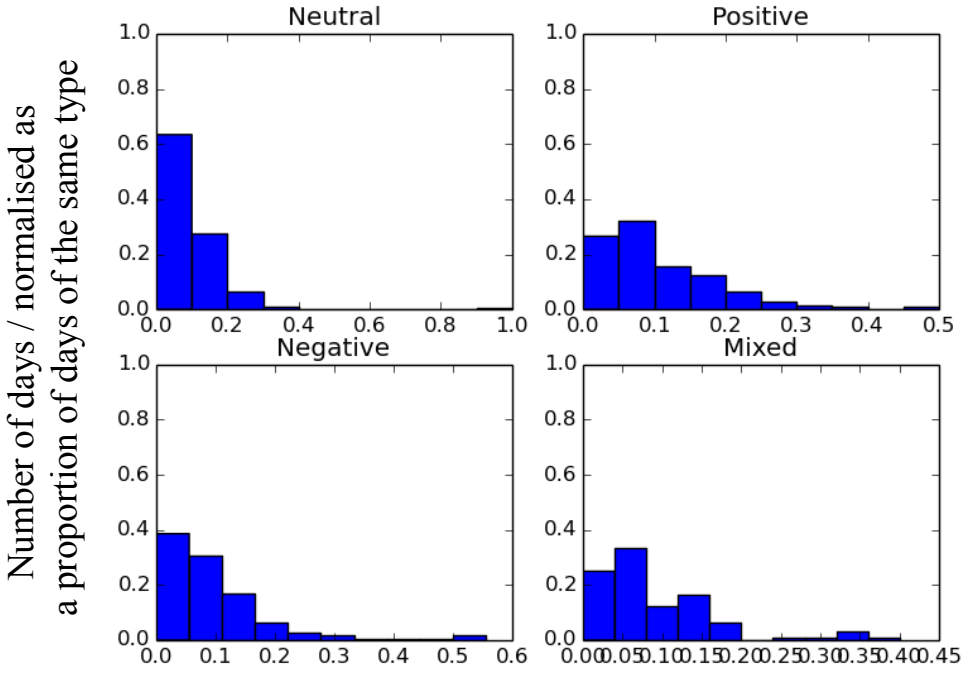
Supplementary Material

Supplementary Figure 1. Timeline of days between 15th May and 13th July 2015 characterised by the type of day for Daesh: positive, negative, neutral or mixed (mixed is represented as both positive and negative with grey shading in between).

A

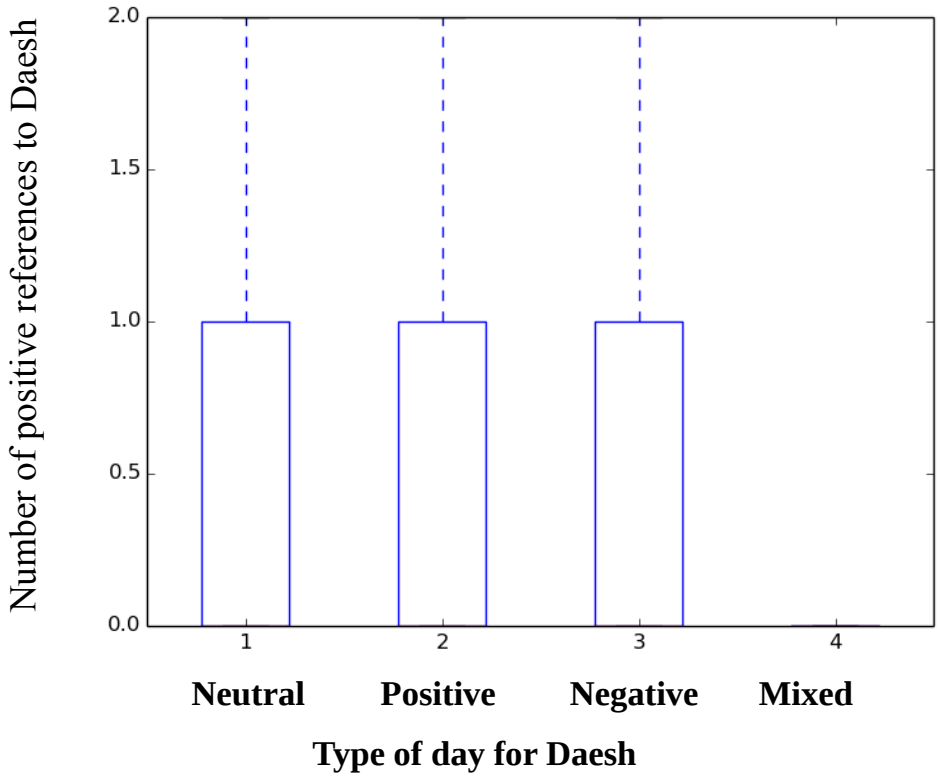


B

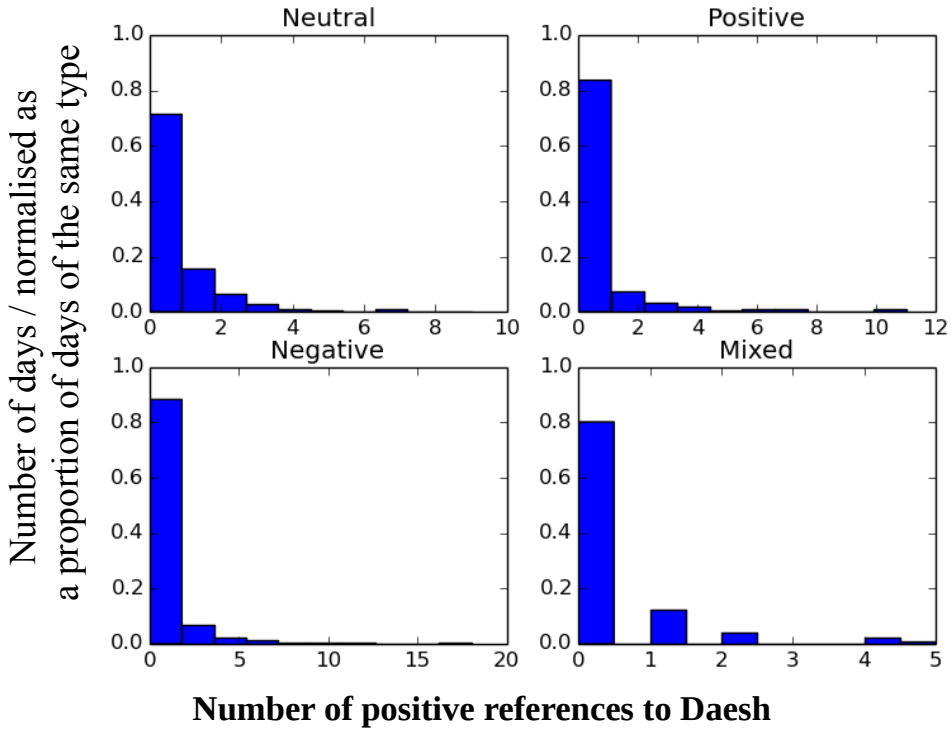


Supplementary Figure 2. Proportion of positive references to Daesh on different types of day (neutral, positive, negative and mixed) is not significantly different. Box and whisker plot (panel A); normalised histogram (panel B). No significant (Bonferroni corrected), Mann Whitney U test p-values.

A



B



Supplementary Figure 3. Number of negative references to Daesh on different types of day (neutral, positive, negative and mixed) is not significantly different. Box and whisker plot (panel A); normalised histogram (panel B). No significant (Bonferroni corrected), Mann Whitney U test p-values.

Supplementary List 1. Entities emerging from a thematic analysis of the tweets.

The following lists groups, entities and individuals mentioned that emerged from thematic analysis:

ISIS/IS/Daesh	Muslim Brotherhood
US/Western Countries	Egyptian Government
Kuffar	Assad (Regime)
Kurds/Peshmerga	Christians
Afghan Government	LGBT+
Muslim prisoners	Turkey
Kashmir/Indian Gov't	Jordan
YPG	The Media
PKK	Nigeria
Shia Muslims	Cameroon
Israel	KSA
AQ (al-Qaeda)	Maghreb Governments
SIC	Hizbollah
Sahwat	SFG
Coalition	HSM
Taliban	Muj (Mujahideen)
Hamas	Ethiopia
Iran	Internal
JaN (Jabhat al Nusra)	JAF
FSA (Free Syrian Army)	

8. Discussion

8.1. Discussion of results

Scientific, computational and big data methods have not been widely applied to tackle social and geopolitical questions. In part, this is because the nature of the data and the questions raise difficulties with the standard methods as natural and computer scientists currently understand them and in part because the technical expertise rarely exists amongst the social science community. This, therefore, presents an opportunity. In this thesis, from a computer and natural scientific viewpoint, I have applied computational and machine methods to novel questions, adapting them to develop solutions to overcome some of the ambiguity, subjectivity and noise problems and from a social scientific point of view, I have applied novel methods to existing problems and areas of study. The overarching conclusion of this thesis is that this can be done reliably—machine and other quantitative, computational methods have a lot to offer to the social sciences—but that caution must be exercised as false positive patterns can be found throughout the data and thus rigourous experimental design is crucial.

In the first chapter, I showed that applying quantitative tools from genetics to large volumes of Twitter data using automated computer scripts revealed novel insights into language evolution, relative to previous analyses of small volumes of transcribed telephone conversations or laboratory experiments. By controlling for sources of variance using the framework of an experiment in genetics, it was possible to eliminate the noise and provide evidence of a convincing and meaningful signal. Not only did this show that a model based on an internal store of word frequencies is consistent with findings (thus enabling novel collaborative work that demonstrated a level at which language inheritance is neutral), but this work underpins the assumption underlying this thesis that differences between and changes over time in, people's and groups' languages

can be indicative of and a proxy for, some of the variation caused by other behaviours such as interactions or communication.

Although the next section of the thesis addressed a more specific phenomenon—resurgent Twitter accounts—my novel approach to finding and analysing it built upon the first finding that Twitter profile language can reveal meaningful, underlying patterns. By applying novel machine methods to thousands of accounts, I was able to characterise a substantially larger set of resurgents than the five previous case studies, thus overturning previous conclusions that suspending accounts is purely disruptive to terrorists. This has obvious policy implications, for the social networks themselves and for legislators deciding how to tackle online radicalisation and propaganda.

Although the approach developed in this thesis to characterise resurgents had a validity in its usefulness at predicting accounts that were subsequently manually verified, in the next section of this thesis I carried out a much more rigourous evaluation of its performance and the theory underlying it. Using a standard machine learning approach, I showed that text-similarity based machine approaches can perform better and quicker, than existing manual annotation. This is not surprising, given that existing methods have led to only five case studies (Stern and Berger, 2015; Berger and Perez, 2016) and a series of acknowledgements of their existence in the limitations sections of papers. As a result of this work, future researchers, in both the field of terrorism and others where social media accounts may be subject to suspension and resurgence (e.g. organised crime or online grooming), will be able to improve the quality of their data with less effort than previously required. These methods could also extend similar research connecting accounts across social networks (Goga *et al.*, 2015; Korula and Lattanzi, 2014; Malhotra, 2013; Vesdapunt and Garcia-Molina, 2014) whether for law enforcement, marketing, or sociological/political analysis reasons. The success and ease of the method also highlights the risks of details about oneself being distributed across multiple accounts. In this work I also investigated the underlying theory behind the text-similarity based approach, showing that the mechanism by which human annotation finds resurgents is independent of expertise or background and is tapping into salient information and features that are also amenable to machine methods.

This work, however, also raises the issue that analysis of social science data, dependent on manual annotation, is inherently circular when the human-coded, test dataset for the development of novel machine methods is hypothesised to perform worse than the novel method. This is the clear limitation with this work, although the small number of positive controls enabled this to be overcome to an extent.

The final section of this thesis addressed a different problem of relating analysis of online language to offline events. It demonstrated that the volume of positive tweets on a given day differs depending on the parity of offline events occurring that day. Furthermore, this was novelly demonstrated internally within the Baqiya family's tweets, showing that previously findings were not exclusively driven by opponents of Daesh being more vocal on specific types of day. The main finding of this section of thesis, however, was a highlight of the issues with computational and discourse methods in this domain. Computational metrics are unable to reliably detect changes in sentiment as we cannot reliably evaluate their results compared to the noise in control periods and discourse analysis methods are too time consuming to generate enough data to detect the subtle changes. One of the aspects that qualitative data can capture is nuance, which quantitative approaches can only capture in so far as they are built in to the questions asked. Future work to further determine how the two methodological traditions can work together to enhance knowledge is important.

8.1.1. Limitations

One of the main reasons for caution in generalising this work is the homophily-based approach to terrorist signals adopted, rather than an *a priori* prescribed definition. As discussed in the literature review, Twitter users rarely (only 13%—Wright *et al.*, 2016) identify a member of a proscribed terrorist organisation and for egotistical and security reasons, people inflate or mask their importance, connections and level of violence. Objectively verifying whether Twitter accounts belong to terrorists is therefore not possible. This scope of this thesis, therefore, is simplified to those who are highly interlinked with other terrorist or extremist accounts—albeit in parts confirmed by inspection that users are extremists. As discussed earlier, association obviously does not make a person a terrorist, but by contributing to the phenomenon of online terrorism, are still worth studying. The principle of homophily (McPherson *et al.*, 2001) also

makes it likely that many of those reciprocally following terrorists would themselves be terrorists or extremists.

Even having established the principle of studying interconnected accounts, snowball sampling methods both limit the ability to reach disjoint groups and exhibit bias towards their seed lists—as discussed in more detail in the discussion and limitations section of *'Chapter 5. Resurgent Insurgents: Quantitative Research Into Jihadists Who Get Suspended but Return on Twitter'* (Wright *et al.*, 2016). Thus, although the workflow was carefully designed to include steps weighting sampling to accounts meeting a threshold of interconnectedness and excluding accounts that were unrepresentative in their popularity, there are limits on generalising our sample to the unofficial, English-speaking, Jihadist, Twitter community and these findings could also benefit from more work with a broader sampling procedure. Although the 10% threshold appeared, via manual inspection of the resultant data, to reliably produce "jihadist-linked" accounts, any threshold will still be arbitrary and relies both on the accuracy of the account included on the jihadist seed list and the assumption that accounts followed by 10% of jihadists are worth sampling and characterising.

In addition, the largely faction and hierarchy free nature of the Baqiya family, outlined in the literature review and throughout the thesis, means that the data is not available to enable us to make statements about the differences between specific terrorist groups, an area of questions important to many terrorism studies researchers.

As well as knowing whether accounts were terrorists, ambiguity and the lack of an objectively annotated data presented a range of other problems that had to be overcome in several studies in the thesis. Although careful thought was given to ensure that an appropriate choice was made at each point, several assumptions and generalisations were made and thus the conclusions of this work are qualified and should be taken in the light of the assumptions made. As discussed further in the limitations section of *'Chapter 7. Bickering Families: an Analysis of the Baqiya Family on Twitter and Offline Daesh Events'*, these assumptions include subgroup assignments, the unusual adaptation of content analysis for comparison with quantitative data, the adaptation of sentiment and text analysis into comparable metrics and the use of a pre-annotation list of offline events.

A completely different question considers the merit of the work. Could government agencies such as the British GCHQ or the American NSA have already covered this work in secret? If such a scenario is possible, was the work still worth carrying out? Several arguments suggest that it is. Firstly, even if these questions have been tackled in secret, the results have not been publicly published. Without the answers, therefore, academics cannot make progress on studying terrorism on Twitter and therefore rediscovering them publicly is important. Secondly, some of the questions are quite complex and may well not have been answered, even in secret. Thirdly, some of the approaches covered in this thesis have come from obscure angles, are therefore even if the questions have been tackled, it is unlikely that it has been through these methods. Next, even though the methodologies and questions tackled here are easily adapted for research in other (non-secret) fields (e.g. marketing, linguistics, horizontal gene transfer, criminology of grooming/gangs/bullying), the questions addressed in this thesis have not been answered. Fifthly, even if the answers are known in secret, science is always stronger with replication and independent discovery. Finally, the secret nature means that it is not possible to know whether it has been carried out and thus, in the absence of any evidence, the benefits of publicly tackling these questions make the gamble worth it.

8.1.2. Future work

The findings in *'Chapter 6. Evaluating Machine and Crowdsourcing Methods for Classifying Pseudoreplicate Terrorist Supporting Accounts on Twitter'* suggest that text-similarity-based machine methods can find pseudoreplicate terrorist accounts better (recall = 95.5%; G mean = 0.977; F score = 0.636) than human annotation of an entire big-data-set (recall = 9.91%; G mean = 0.315; F score = 0.139; performance evaluated against positive controls). There may, however, be scope to improve this performance by including more features, such as the analysis of tweets. Furthermore, although the accounts matched by the model were significantly less likely to have screenshots than the average terrorist supporting account, including imagery analysis into the machine method could still be of use in some cases.

The findings in *'Chapter 7. Bickering Families: an Analysis of the Baqiya Family on Twitter and Offline Daesh Events'*, that sentiment on Twitter is associated with the parity of offline events, has potential for development, but there are many unanswered questions. It is not known, for example, whether the sentiment is driven by media reports of events, or whether the Baqiya family has direct access to the events as they occur. If the latter were to be the case, then analysing the sentiment of the Baqiya family might provide information on events before the media reports them. If not, then detecting changes from the Baqiya family's normal response to media reports might also be informative.

8.2. Conclusion

Twitter, social media and big data promises much in terms of terrorist signals amenable to analysis. Together, the work in this thesis shows that computational analysis of big data enables tuning in to subtle signals and sometimes reveals conclusions that contradict less developed research. Control noise, however, often contains as many patterns and, thus, future studies should pay particular attention to their methodologies when using noisy, subjective, social media data.

9. Bibliography

Abc.net, 2013. Indonesian plotted on Facebook to attack Myanmar embassy. *abc.net,* 6 November [online]. Available at: http://www.abc.net.au/news/2013-11-06/an-indonesian-plotted-on-facebook-to-attack-myanmar-embassy/5074782

Agence France Presse, 2013. Al Qaeda Offshoot Joins Twitter, AQIM Threatens To Execute French Hostages. *The Huffington Post.* [online]. Available at: http://www.huffingtonpost.com/2013/03/28/al-qaeda-in-the-islamic-maghreb-joins-twitter_n_2972907.html [Accessed 13 February 2014]

Agence France Presse, 2014. Twitter suspends account of Hamas armed wing. *Yahoo News.* 23 January [online]. Available at: http://news.yahoo.com/twitter-suspends-account-hamas-armed-wing-141443942.html [Accessed 13 February 2014]

Aldrich, R.J., 2011. *GCHQ: The Uncensored Story of Britain's Most Secret Intelligence Agency*. London: Harper*Press*

Amarasingam, A., 2015. What Twitter Really Means For Islamic State Supporters. *War on the Rocks.* December 30.

Andrew, C., Gordievsky, O., 1990. *KGB: The Inside Story of its Foreign Operations from Lenin to Gorbachev*. London: Hodder and Stoughton

Arthur, C., 2014. Taking down Isis material from Twitter or YouTube not as clear cut as it seems. *The Guardian*. June 23.

Barnett, D., 2013. Are you looking at an official Shabaab Twitter account. *Threat Matrix: A Blog Of The Long War Journal.* [online] Available at: http://www.longwarjournal.org/threat-matrix/archives/2013/09/are_you_looking_at_an_official.php [Accessed 13 February 2014]

BBC, 2015. Charlie Hebdo attack: Three days of terror. BBC News. January 14 [online]. Available at: http://www.bbc.co.uk/news/world-europe-30708237

Bennett, C.M., Baird, A.A., Miller, M.B., Wolford, G.L., 2011. Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon: An Argument For Proper Multiple Comparisons Correction. *Journal of Serendipitous and Unexpected Results*. **1**(1): 1-5.

Berger, J.M., Morgan, J., 2015. The ISIS Twitter Census: Defining and describing the population of ISIS supporters on Twitter. *The Brookings Institution.*

Berger, J.M., Perez, H., 2016. The Islamic State's Diminishing Returns on Twitter: How suspensions are limiting the social networks of English-speaking ISIS supporters [Occasional Paper]. *George Washington Program on Extremism.*

Birky, C.W., 1995. Uniparental inheritance of mitochondrial and chloroplast genes: mechanisms and evolution. *PNAS*. **92**(25), 11331–11338

Black Standard, (n.d.). In *Wikipedia.* Retrieved Aug 23, 2016, from https://en.wikipedia.org/wiki/Black_Standard

Bloomfield, L., 1933. *Language.* University of Chicago Press.

Blythe, R.A., 2012. Neutral evolution: A null model for language dynamics. *Advances in Complex Systems*. **15**(03n04):1150015

Bodine-Baron, E., Helmus, T.C., Magnuson, M., Winkelman, Z., 2016. *Examining ISIS Support and Opposition Networks on Twitter.* Santa Monica, CA: The RAND Corporation. Available at: http://www.rand.org/pubs/research_reports/RR1328.html [Accessed 17 August 2016]

Bonduriansky, R., Day, T., 2009. Nongenetic Inheritance and Its Evolutionary Implications. *Annual Review of Ecology, Evolution and Systematics*. **40**(1):103-125

Bonduriansky, R., 2012. Rethinking heredity, again. *Trends in ecology and evolution*. **27**(6):330-6

Bonferroni, C.E., 1936. Teoria statistica delle classi e calcolo delle probabilita. *Libreria internazionale Seeber.*

Bonin, F., De Looze, C., Ghosh, S., *et al.*, 2013. Investigating fine temporal dynamics of prosodic and lexical accommodation. *INTERSPEECH 2013.* 539-543

Branigan, H., Pickering, M., Pearson, J., *et al.*, 2011. The role of beliefs in lexical alignment: evidence from dialogs with humans and computers.. *Cognition.* **121**(1):41-57

Bray, J.R., Curtis, J.T., 1957. An ordination of the upland forest communities of southern Wisconsin. *Ecological monographs.* **27**(4):325–349.

Brynielsson, J., Horndahl, A., Johansson, F., *et al.*, 2012. Analysis of Weak Signals for Detecting Lone Wolf Terrorists. *2012 European Intelligence and Security Informatics Conference*. 197-204. DOI 10.1109/EISIC.2012.20

Brennan, S.E., 1996. Lexical entrainment in spontaneous dialog. *Proceedings of ISSD*. **96**:41-44

Bryden, J., Funk, S., Geard, N., Bullock, S., Jansen, V.A.A., 2011. Stability in flux: community structure in dynamic networks. *Journal of the Royal Society, Interface.* **8**(60):1031-40.

Bryden, J., Funk, S., Jansen, V.A.A., 2013. Word usage mirrors community structure in the online social network Twitter. *EPJ Data Science.* **2**(3)

Bunzel, C., 2016. "Come Back to Twitter": A Jihadi Warning Against Telegram. *Jihadica*. 18 July [online]. Available at: http://www.jihadica.com/come-back-to-twitter/ [Accessed 21 August 2016]

Burrell, J., 2016. How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society.* **3**(1):1-12 DOI: 10.1177/2053951715622512

Carlile, Lord of Berriew Q.C., 2007. The Definition of Terrorism. (Cmnd. 7052). *UK Parliamentary Report.* London.

Carone, B.R., Fauquier, L., Habib, N., *et al.* 2010. Paternally induced transgenerational environmental reprogramming of metabolic gene expression in mammals. *Cell*. **143**(7):1084-96

Caruana, R., Lou, Y., Gehrke, J., *et al.*, 2015. Intelligible Models for HealthCare. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15.* 1721-1730. DOI: http://dx.doi.org/10.1145/2783258.2788613

Castellano, C., Fortunato, S., Loreto, V., 2009. Statistical physics of social dynamics. *Reviews of Modern Physics*. **81**(2):591-646

Cavalli-Sforza, L.L., Feldman, M.W., 1981. *Cultural transmission and evolution: a quantitative approach*. Princeton University Press, No. 16

Chee, M.W., Hon, N.H., Caplan, D., Lee, H.L., Goh, J., 2002. Frequency of concrete words modulates prefrontal activation during semantic judgments. *Neuroimage*. **16**(1):259–268.

Chater, N., Christiansen, M.H., 2010. Language Acquisition Meets Language Evolution. *Cognitive Science*. **34**(7):1131-1157

Chatfield, A.T., Reddick, C.G., Brajawidagda, U., 2015. Tweeting propaganda, radicalization and recruitment: Islamic state supporters multi-sided twitter networks. *Proc 16th Annual International Conference on Digital Government Research*. pp.239-49 DOI:10.1145/2757401.2757408

Christopherson, L., 2011. Can u help me plz?? Cyberlanguage accommodation in virtual reference conversations. *Proceedings of the American Society for Information Science and Technology.* **48**(1): 1-9

Church, K.W., 2000. Empirical Estimates of Adaptation : The chance of Two Noriegas is closer to p/2 than p2. *Proceedings of the 18th International Conference on Computational Linguistics*. C00-1027

Clarke, L., 2009. Why has Defining Terrorism Proved so Difficult? *e-International Relations,* [online]. Available at: http://www.e-ir.info/2009/05/14/why-has-defining-terrorism-proved-so-difficult/

Cooper, H.H.A., 1978. Psychopath As Terrorist. Legal Medical Quarterly. **2**(4): 253-262

Cormack, G., Lynam, T., 2005. TREC 2005 Spam Track Overview. *Sixteenth Text Retrieval Conference (TREC 2007).* 1-9

Corner, E., Gill, P., 2015. A False Dichotomy? Mental Illness and Lone-Actor Terrorism. *Law and Human Behaviour*. **39**(1): 23–34. doi: 10.1037/lhb0000102

Council of the European Union, 2002. Council Framework Decision of 13 June 2002 on Combating Terrorism. *Official Journal of the European Communities.* 164: 3-7

Croft, W., 2000. *Explaining Language Change: an Evolutionary Approach*. Pearson Education

Croft, W., 2006. *The Relevance of an Evolutionary Model to Historical Linguistics*. In: Competing Models of Linguistic Change: Evolution and beyond (Thomsen, O.N., ed.). doi: 10.1075/cilt.279.08cro

Cronin, A.K., Ludes, J.M., (eds.), 2004. *Attacking terrorism: Elements of a grand strategy.* Washington DC: Georgetown University Press.

Danchin, É., Wagner, R.H., 2010. Inclusive heritability: combining genetic and non-genetic information to study animal behavior and culture. *Oikos*. **119**(2): 210-8

Danchin, É., Charmantier, A., Champagne, F.A., *et al.*, 2011. Beyond DNA: integrating inclusive inheritance into an extended theory of evolution. *Nature Reviews Genetics.* **12**(7): 475-86.

Danescu-Niculescu-Mizil, C., Gamon, M., Dumais, S., 2011. Mark My Words! Linguistic Style Accommodation in Social Media. *World Wide Web Conference*. ACM 978-1-4503-0632-4/11/03

Darwin, C., 1859. *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. London: Murray.

De Looze, C., Oertel, C., Rauzy, S., Campbell, N., 2011. Measuring dynamics of mimicry. *ICPhS XVII.* 1294-1297

De Looze, C., Scherer, S., Vaughan, B., Campbell, N., 2014. Investigating automatic measurements of prosodic accommodation and its dynamics in social interaction. *Speech Communication.* **58**:11-34 DOI: 10.1016/j.specom.2013.10.002

Definitions of terrorism: United States, (n.d.). In *Wikipedia.* Retrieved November 1, 2013, from http://en.wikipedia.org/wiki/Definitions_of_terrorism#United_States

di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., Rizzolatti, G., 1992. Understanding motor events: a neurophysiological study. *Experimental Brain Research.* **91**(1): 176-180

Diakopoulos, N., 2014. Algorithmic Accountability Reporting: On the Investigation of Black Boxes. *Tow Center for Digital Journalism*.

Dodd, B., Holm, A., Hua, Z., Crosbie, S., 2003. Phonological development: a normative study of British English-speaking children. *Clinical Linguistics & Phonetics*. **17**(8): 617–643

Dodd, V., Halliday, J., Watt, N., 2015. UK police on highest ever terror alert after Belgian arrests. The Guardian. January 15 [online]. Available at: https://www.theguardian.com/uk-news/2015/jan/16/uk-police-high-terror-alert-severe-threat

Domingos, P., 2012. A few useful things to know about machine learning. *Communications of the ACM.* **55**(10): 78

Dominiczak, P., 2014. Theresa May: Britain is facing greatest terror threat of its history. The Telegraph. November 24 [online]. Available at: http://www.telegraph.co.uk/news/uknews/terrorism-in-the-uk/11249614/Theresa-May-Britain-is-facing-greatest-terror-threat-of-its-history.html

Dunn, M., Terrill, A., Reesink, G., Foley, R.A., Levinson, S.C., 2005. Structural phylogenetics and the reconstruction of ancient language history. *Science.* **309**(5743): 2072-2075

Dvorin, T., 2014. Cyber Victory: Hamas Twitter Accounts Suspended. *Israel National News.* 19 January [online]. Available at: http://www.israelnationalnews.com/News/News.aspx/176449#.Uvy2EefftHY [Accessed 13 February 2014]

Eisenstein, J., O'Connor, B., Smith, N.A., Xing, E.P., 2012. Mapping the geographical diffusion of new words. *arXiv preprint*. arXiv:1210.5268

Eisenstein, J., O'Connor, B., Smith, N.A., Xing, E.P., 2014. Diffusion of lexical change in social media. *PLoS ONE.* **9**(11): e113114

Fairclough, N., 2003. *Analysing Discourse: Textual Analysis for Social Research*. London: Routledge

Fairclough, N., 2010. *Critical discourse analysis: the critical study of language*. London: Routledge

Falconer, D.S., Mackay, T.F.C., 1995. *Introduction to Quantitative Genetics* (4th ed.). Longman

Ferrara, E., Wang, W., Varol, O., *et al.*, 2016. Predicting online extremism, content adopters and interaction reciprocity. *arXiv preprint*. arXiv:1605.00659

Fisher, M., 2012. Qaeda-linked Syrian rebel group is feuding with WordPress on Twitter. *The Washington Post.* 13 December [online]. Available at: http://www.washingtonpost.com/blogs/worldviews/wp/2012/12/13/qaeda-linked-syrian-rebel-group-is-feuding-with-wordpress-on-twitter/ [Accessed 13 February 2014]

Fisher, A., 2015. Swarmcast: How Jihadist Networks Maintain a Persistent Online Presence. *Perspectives on terrorism*. **9**(3)

Gallagher, E.D., 1999. *COMPAH Documentation*. Boston, MA: Environmental, Coastal & Ocean Sciences Department, University of Massachusetts at Boston, 1-59 [online] Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.9.1334&rep=rep1&type=pdf [Accessed 18 August 2016]

Gallois, C., Ogay, T., Giles, H., 2005. Communication accommodation theory: A look back and a look ahead. In: *Theorizing About Intercultural Communication*, ed. Gudykunst, W.B. California: Sage, pp.121-148

Ganor, B., 2010. Defining Terrorism - Is One Man's Terrorist Another Man's Freedom Fighter? *International Institute for Counter-Terrrorism,* [online] Available at: http://www.ict.org.il/ResearchPublications/tabid/64/Articlsid/432/Default.aspx

Garner, G., 2010. Case Studies in Exploiting Terrorist Group Divisions with Disinformation and Divisive/Black Propaganda. *Journal of Terrorism Research*. **1**(1). DOI: http://dx.doi.org/10.15664/jtr.164

Gertz, B., 2013. Al Qaeda opens first official Twitter account. *The Washington Times.* [online]. Available at: http://www.washingtontimes.com/news/2013/sep/27/al-qaeda-opens-first-official-twitter-account/?page=all [Accessed 13 February 2014]

Ghajar-Khosravi, S., Kwantes, P., Derbentseva, N., Huey, L., 2016. Quantifying Salient Concepts Discussed in Social Media Content: A Case Study using Twitter Content Written by Radicalized Youth. *Journal of Terrorism Research*. **7**(2): 79–90. DOI: http://dx.doi.org/10.15664/jtr.1241

Gillespie, T., 2012. The relevance of algorithms [Forthcoming] In: *Media Technologies: Essays on Communication, Materiality and Society,* ed. Gillespie, T., Boczkowski, P., *Foot, K.* Cambridge, MA: MIT Press

Gladstone, R., 2015. Twitter Says It Suspended 10,000 ISIS-Linked Accounts in One Day. *The New York Times*. April 9.

Goga, O., Loiseau, P., Sommer, R., Teixeira, R., Gummadi, K.P., 2015. On the Reliability of Profile Matching Across Large Online Social Networks. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* 1799-1808

Gomaa, W., 2013. A Survey of Text Similarity Approaches. *International Journal of Computer Applications.* **68**(13): 13-18

Goodman, L.A., 1961. Snowball Sampling. *Ann Math Stat.* **32**: 148–170

Gorman, R., 2013. Terrorist group al-Qaeda's Twitter account has been suspended after only five days. *The Daily Mail.* 29 September. [online]. Available at: http://www.dailymail.co.uk/news/article-2438135/Terrorist-group-al-Qaedas-Twitter-account-suspended-days.html [Accessed 13 February 2014]

Gray, R.D., Drummond, A.J., Greenhill, S.J., 2009. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science*. **323**(5913): 479-483

Greenberg J, 2015. Why Facebook and Twitter Can't Just Wipe Out ISIS Online. *WIRED: Business* [newspaper on the Internet]. Nov 21 [cited 2016 Apr 22]. Available at: http://www.wired.com/2015/11/facebook-and-twitter-face-tough-choices-as-isis-exploits-social-media/

Greene, K.J., 2015. ISIS: Trends in Terrorist Media and Propaganda. *International Studies Capstone Research Papers*. Paper 3.

Grieve, J., Nini, A., Guo, D., 2016. Analyzing lexical emergence in Modern American English online. *English Language and Linguistics.* 1-29. doi:10.1017/S1360674316000113

Gurajala, 2015. *Proceedings of the 2015 International Conference on Social Media & Society - SMSociety '15.*

Hall, B. K., Hallgrímsson, B., (Eds.), 2008. *Strickberger's Evolution* (4th ed.). Jones & Bartlett.

Hamm, M., Spaaj, R., 2015. *Lone Wolf Terrorism in America: Using Knowledge of Radicalization Pathways to Forge Prevention Strategies.*

Hardaker, C., 2013. "Uh. . . . not to be nitpicky,,,,,but…the past tense of drag is dragged, not drug.": An overview of trolling strategies. *Journal of Language Aggression and Conflict*. 1(1): 57-86. DOI: 10.1075/jlac.1.1.04har

Harris, C., Srinivasan, P., 2014. Hybrid Crowd-Machine Methods as Alternatives to Pooling and Expert Judgments. *Information Retrieval Technology, Airs 2014, LNCS 8870.* 60-72

Hauser, M.D., Chomsky, N., Fitch, W.T., 2002. The Faculty of Language: What Is It, Who Has It and How Did It Evolve? *Science*. **298**(5598): 1569-1579. DOI: 10.1126/science.298.5598.1569

Hemphill, L., Otterbacher, J., 2012. Learning the Lingo ? Gender , Prestige and Linguistic Adaptation in Review Communities. *2012 ACM Conference on Computer Supported Cooperative Work – CSCW'12 Session: Forums Online.*305-314

Hoffman, B., 2004. *Inside terrorism*. Columbia University Press.

Home Office, (n.d.). PROSCRIBED TERRORIST ORGANISATIONS. Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/538297/20160715-Proscription-website-update.pdf [Accessed 23 August 2016]

Horgan, J., Braddock, K., (eds.), 2012. *Terrorism studies: a reader.* Oxford: Routledge.

Howden, D., 2013. Terror in Nairobi: the full story behind al-Shabaab's mall attack. The Guardian. October 4 [online]. Available at: https://www.theguardian.com/world/2013/oct/04/westgate-mall-attacks-kenya

Hudson, J., 2012. The Most Infamous Terrorists on Twitter. *The Wire.* 2 January. [online]. Available at: http://www.thewire.com/global/2012/01/most-infamous-terrorists-twitter/46852/ [Accessed 13 February 2014]

Huey, L., Witmer, E., 2016. #IS_Fangirl: Exploring a New Role for Women in Terrorism. *Journal of Terrorism Research*. **7**(1): 1-10. DOI: http://dx.doi.org/10.15664/jtr.1211

Hurlbert, S.H., 1984. Pseudoreplication and the Design of Ecological Field Experiments. *Ecological Monographs*. **54**(2): 187-211. DOI: 10.2307/1942661

Iacoboni, M., Woods, R.P., Brass, M., *et al.*, 1999. Cortical Mechanisms of Human Imitation. *Science*. **286**(2526): 2526-8. DOI: 10.1126/science.286.5449.2526

IBM Watson. *AlchemyAPI [computer software].* Available at: http://www.alchemyapi.com/ [Accessed January 2016]

ITV Report, 2013. MI6 chief: UK's enemies 'rubbing their hands with glee' over Snowden leaks. *ITV News.* November 7 [online]. Available at: http://www.itv.com/news/2013-11-07/mi6-chief-uks-enemies-rubbing-their-hands-with-glee-over-snowden-leaks/

ITV Report, 2015a. Al-Qaida terrorists planning 'mass casualty attacks' against the West, MI5 boss warns. *ITV News.* January 8 [online]. Available at: http://www.itv.com/news/2015-01-08/al-qaida-terrorists-planning-mass-casualty-attacks-against-the-west-m15-boss-warns/

ITV Report, 2015b. David Cameron says terror threat to UK is his 'greatest concern' as he renews calls for 'snooper's charter'. *ITV News.* January 11 [online]. Available at: http://www.itv.com/news/2015-01-11/david-cameron-terror-threat-to-uk-is-my-greatest-concern/

Iwata, T., Watanabe, S., 2013. Influence relation estimation based on lexical entrainment in conversation. *Speech Communication.* **55**(2): 329-339

Jablonka, E., Lamb, M.J., 1995. *Epigenetic Inheritance and Evolution.* Oxford, UK: Oxford Univ. Press.

Jablonka, E., Lamb, M.J., 2005. *Evolution in Four Dimensions*. Cambridge, MA:MIT Press

Jimenez-Chillaron, J.C., Isganaitis, E., Charalambous, M., *et al.* 2009. Intergenerational Transmission of Glucose Intolerance and Obesity by In Utero Undernutrition in Mice. *Diabetes*. **58**(2): 460–468

Johnston, P., 2015. Laws against 'extremism' risk criminalising us all. *The Telegraph.* September 28 [online]. Available at: http://www.telegraph.co.uk/news/uknews/law-and-order/11897355/Laws-against-extremism-risk-criminalising-us-all.html [Accessed 23 August 2016]

Jones, S., Ahmed, M., 2014. Tech groups aid terror, says UK spy chief. *Financial Times*. November 3 [online]. Available at: http://www.ft.com/cms/s/2/4a35c0b2-636e-11e4-9a79-00144feabdc0.html [Accessed 18 August 2016]

Jucks, R., Becker, B., Bromme, R., 2008. Lexical Entrainment in Written Discourse: Is Experts' Word Use Adapted to the Addressee? *Discourse Processes*. **45**(6): 497-518

Kim, K.H.S., Relkin, N.R., Lee, K.M., Hirsch, J., 1997. Distinct cortical areas associated with native and second languages. *Nature*. **388**(6638): 171–174.

Kirby, S., Cornish, H., Smith, K., 2008. Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *PNAS*. **105**(31): 10681-6

Kirby, S., Griffiths, T., Smith, K., 2014. Iterated learning and the evolution of language. *Current Opinion in Neurobiology*. **28**: 108-114

Klausen, J., Barbieri, E., Reichlin-Melnick, A., Zelin, A.Y., 2012. The YouTube Jihadists: A Social Network Analysis of Al- Muhajiroun' s Propaganda Campaign. *Perspectives on Terrorism*. **6**(1):36-53

Klausen, J., 2015. Tweeting the Jihad: Social Media Networks of Western Foreign Fighters in Syria and Iraq. *Studies in Conflict & Terrorism*. **38**:1–22. DOI: 10.1080/1057610X.2014.974948

Korevaar, T.I.M., Muetzel, R., Medici, M., *et al.*, 2016. Association of maternal thyroid function during early pregnancy with offspring IQ and brain morphology in childhood: a population-based prospective cohort study. *The Lancet Diabetes & Endocrinology*. **4**(1): 35–43

Korula, N., Lattanzi, S., 2014. An efficient reconciliation algorithm for social networks. *Proceedings of the VLDB Endowment*. **7**(5): 377-388

Laqueur, W., 2012. *A History of Terrorism*. New Jersey: Transaction Publishers

Leavitt, A., 2014. From #FollowFriday to YOLO: Exploring the Cultural Salience of Twitter Memes. In: Weller, Bruns, Burgess, Mahrt, Puschmann. ed. *Twitter and Society*. New York: Peter Lang. pp.137-154.

Lester, D., Yang, B., Lindsay, M., 2004. Suicide Bombers: Are Psychological Profiles Possible? *Studies in Conflict & Terrorism*. **27**: 283–295. DOI: 10.1080/10576100490461033

Levy, R., 2014. ISIS Tries to Outwit Social Networks. *Vocativ*. June 17.

Lieberman, E., Michel, J.B., Jackson, J., Tang, T., Nowak, M.A., 2007. Quantifying the evolutionary dynamics of language. *Nature*. **449**(7163): 713–716

Lipton, Z.C., 2016. The Mythos of Model Interpretability. *arXiv preprint*. arXiv:1606.03490v1

List of terrorist incidents linked to ISIL, (n.d.). In *Wikipedia*. Retrieved May 25, 2016, from https://en.wikipedia.org/wiki/List_of_terrorist_incidents_linked_to_ISIL#2016

Lo, S., Chiong, R., Cornforth, D., 2015. Using support vector machine ensembles for target audience classification on Twitter. *PLoS ONE.* **10**(4): 1-20 DOI: 10.1371/journal.pone.0122855

Longa, V.M., 2013. The evolution of the Faculty of Language from a Chomskyan perspective: bridging linguistics and biology. *Journal of Anthropological Sciences*. 91: 1-48. doi 10.4436/JASS.91011

Lynch, M., Freelon, D., Aday, S., 2014. *BLOGS AND BULLETS III : SYRIA'S SOCIALLY MEDIATED CIVIL WAR.* Washington: United States Institute of Peace

Magdy, W., Darwish, K., Weber, I., 2015. #FailedRevolutions: Using Twitter to Study the Antecedents of ISIS Support. *arXiv preprint.* arXiv:1503.02401v1

Magdy, W., Darwish, K., Weber, I., 2016. #FailedRevolutions: Using Twitter to study the antecedents of ISIS support. First Monday. **21**(2): 1-15

Mahmood, S., 2012. Online social networks: Te overt and covert communication channels for terrorists and beyond. *2012 IEEE Conference on Technologies for Homeland Security (HST).*

Mairesse, F., Walker, M.A., Mehl, M.R., Moore, R.K., 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research.* **30**(1): 457–500

Malhotra, A., 2013. Studying User Footprints in Different Online Social Networks. *arXiv preprint*. arXiv:1301.6870v1

McCarthy, N., 2015. Belgium Is The EU 'Capital' For Foreign Fighters. *statista*. November 18, [online]. Available at: https://www.statista.com/chart/4024/belgium-is-the-eu-capital-for-foreign-fighters/ [Accessed 23 August 2016]

McPherson, M., Smith-Lovin, L., Cook, J.M., 2001. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology.* **27**: 415-44.

Meisels, T., 2008. *The trouble with terror: Liberty, Security and the Response to Terrorism*. Cambridge, UK: Cambridge University Press.

Melle, I., 2013. The Breivik case and what psychiatrists can learn from it. *World Psychiatry*. **12**(1): 16–21. doi: 10.1002/wps.20002

Mesoudi, A., McElligott, A.G., Adger, D., 2011. Introduction: Integrating Genetic and Cultural Evolutionary Approaches to Language. *Human Biology*. **83**(2)

Mesoudi, A., Blanchet, S., Charmantier, A., *et al.* 2013. Is Non-genetic Inheritance Just a Proximate Mechanism? A Corroboration of the Extended Evolutionary Synthesis. *Biological Theory*. **7**(3): 189-95.

Meisels, T., 2008. *The trouble with terror: Liberty, Security and the Response to Terrorism*. Cambridge, UK: Cambridge University Press.

Miller, G.A., 1956. Te magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review.* **63**(2): 81–97.

Miller, N., 2015. Global terror 'family' Baqiya a growing concern for security. *The Sydney Morning Herald*. 3rd October [online]. Available at: http://www.smh.com.au/world/global-terror-family-baqiya-a-growing-concern-for-security-20151002-gk0fq5.html [Accessed 13 January 2016]

Mitchell, 2012. *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL).*

Mohamed, H., 2013. Al-Shabab say they are back on Twitter. *ALJAZEERA.* [online]. Available at: http://www.aljazeera.com/news/africa/2013/12/al-shabab-claim-they-are-back-twitter-2013121610453327578.html [Accessed 13 February 2014]

Mohammad, S.M., Yang, T., 2013. Tracking Sentiment in Mail: How Genders Differ on Emotional Axes. *arXiv preprint*. arXiv:1309.6347v1

Moriarty, B., 2015. Defeating ISIS on Twitter. *Technology Science*. 2015092904.

Morris, N., 2014. MPs' Iraq vote: Cameron warns Isis air strikes not enough to defeat 'bunch of psychopathic terrorists'. *Independent*. 26 September [online]. Available at: http://www.independent.co.uk/news/uk/politics/mps-iraq-vote-cameron-warns-isis-air-strikes-not-enough-to-defeat-bunch-of-psychopathic-terrorists-9757715.html [Accessed 21 August 2016]

Moskalenko, S., McCauley, C., 2011. The psychology of lone-wolf terrorism. *Counselling Psychology Quarterly.* **24**(2): 115-126. doi: 10.1080/09515070.2011.581835

National Consortium for the Study of Terrorism and Responses to Terrorism (START), 2016. Global Terrorism Database [Data file]. Available at: https://www.start.umd.edu/gtd [Accessed 17 August 2016]

Naylor, F., 2011. Emily Wilding Davison Martyrs or Firebrand? *Higher: The magazine for the alumni of Royal Holloway and Bedford*. **15**(Winter) Available at: https://www.royalholloway.ac.uk/alumni/documents/pdf/higher/higher15.pdf [Accessed 17 August 2016]

Nenkova, A., Gravano, A., Hirschberg, J., 2008. High Frequency Word Entrainment in Spoken Dialogue. *Proceedings of ACL-08: HLT, Short Papers (Companion Volume)*. 169-172

Nerbonne, J., 2010. Measuring the diffusion of linguistic change. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences.* **365**: 3821-3828

Ng, S.F., Lin, R.C., Laybutt, D.R., *et al.* 2010. Chronic high-fat diet in fathers programs β-cell dysfunction in female rat offspring. *Nature*. **467**(7318): 963-6

Nowak, M.A., Plotkin, J.B., Jansen, V.A.A., 2000. The evolution of syntactic communication. *Nature*. **404**(6777): 495-8

Nowak, M.A., Komarova, N.L., Niyogi, P., 2001. Evolution of universal grammar. *Science*. **291**(5501): 114-118

Nowak, M.A., Komarova, N.L., Niyogi, P., 2002. Computational and evolutionary aspects of language. *Nature*. **417**(6889): 611-617

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

O'Connor, A., 2004. Understanding Who Becomes Terrorists. *Journal of Young Investigators.* **11**(3).

Office of the Directory of National Intelligence, 2015. Bin Laden's Bookshelf. [online] Available at: https://www.dni.gov/index.php/resources/bin-laden-bookshelf [Accessed 18 August 2016]

Pagel, M., Atkinson, Q.D., Meade, A., 2007. Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature.* **449**(7163): 717–720.

Pagel, M., 2009. Human language as a culturally transmitted replicator. *Nature Reviews Genetics*. **10**(6): 405-15

Pardo, J.S., 2006. On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America.* **119**(4): 2382-2393

Pardo, J.S., Gibbons, R., Suppes, A., Krauss, R., M., 2012. Phonetic convergence in college roommates. *Journal of Phonetics.* **40**:190-197. doi:10.1016/j.wocn.2011.10.001

Pasquale, F., 2009. Assessing algorithmic authority. [blog] Madisonian. Available at: http://madisonian.net/2009/11/18/assessing-algorithmic-authority/ [Accessed 11 July 2016]

Pennebaker, J.W., Mehl, M.R., Niederhoffer, K.G., 2003. PSYCHOLOGICAL ASPECTS OF NATURAL LANGUAGE USE: Our Words, Our Selves. *Annu. Rev. Psych*. **54**: 547-77. doi: 10.1146/annurev.psych.54.101601.145041

Popper, K., 1959. *The Logic of Scientific Discovery.* London: Routledge

Preoțiuc-Pietro, D., Volkeva, V., Lampos, V., Bachrach, Y., Aletras, N., 2015. Studying User Income through Language, Behaviour and Affect in Social Media. *PLoS ONE.* **10**(9): e0138717. doi:10.1371/ journal.pone.0138717

Radford, E.J., Ito, M., Shi, H., 2014. In utero undernourishment perturbs the adult sperm methylome and intergenerational metabolism. *Science*. **345**(6198): 1255903

Rae, J., 2012. Will it Ever be Possible to Profile the Terrorist? *Journal of Terrorism Research*. **3**(2).

Rasch, W., 1979. Psychological Dimensions of Political Terrorism in the Federal Republic of Germany. *International Journal of Law and Psychiatry*. **2**: 79-85

Rowe, M., Saif, H., 2016. Mining Pro-ISIS Radicalisation Signals from Social Media Users. *Association for the Advancement of Artificial Intelligence.*

Ryan, L., 2014. Al-Qaida and ISIS Use Twitter Differently: Here's How and Why. *National Journal*. October 9.

Rykiel, E.,1996. Testing ecological models: The meaning of validation. *Ecological Modelling.* **90**(3): 229-244

Sageman, M., 2004. *Understanding Terror Networks.* Philadelphia: University of Pennsylvania Press.

Salton, G., McGill, M.J., 1983. *Introduction to modern information retrieval.* McGraw-Hill.

Sarasso, S., *et al.*, 2014. Plastic Changes Following Imitation-Based Speech and Language Therapy for Aphasia A High-Density Sleep EEG Study. *Neurorehabilitation and neural repair*. **28**(2):129–138.

Schmid, A.P., 2011. *The Routledge Handbook of Terrorism Research.* Abingdon: Taylor & Francis

Schneider, F., 2013. How To Do A Discourse Analysis: A toolbox for analysing political texts. *Politics East Asia.*

Schneier, B., 1996. *Applied Cryptography: Protocols, Algorithms and Source Code in C* (2nd ed). New York: John Wiley & Sons

Seaver, N., 2014. Knowing algorithms! *Media in Transition 8.* 1-12

Shariatmadari, D., 2014. Why there's no such thing as Islamic State. *The Guardian.* October 1 [online]. Available at: https://www.theguardian.com/commentisfree/2014/oct/01/islamic-state-language-isis?CMP=twt_gu [Accessed 17 August 2016]

Shirky, C., 2009. A Speculative Post on the Idea of Algorithmic Authority. [Blog] *Clay Shirky*. Available at: http://www.shirky.com/weblog/2009/11/a-speculative-post-on-the-idea-of-algorithmic-authority/ [Accessed 11 July 2016]

Silke, A., (ed.), 2009. *Terrorists, Victims and Society: Psychological Perspectives on Terrorism and its Consequences.* Chichester: John Wiley & Sons Ltd.

Smucker, M.D., Kazai, G., Lease, M., 2012. Overview of the TREC 2012 Crowdsourcing Track. *Proceedings of the 21st NIST text retrieval conference (TREC-2012).*

Sokolova, M., Japkowicz, N., Szpakowicz, S., 2006. Beyond accuracy, F-Score and ROC: A family of discriminant measures for performance evaluation. *Advances in Artificial Intelligence, AI 2006, LNAI 4304.* 4304:1015-1021

Standing, L., 1973. Learning 10,000 pictures. *Q J Exp Psychol*. **25**:207-22.

Steels, L., Kaplan, F., 2002. Aibo's first words: The social learning of language and meaning. *Evolution of communication*. **4**(1):3-32

Steinhauser, N., Campbell, G., Taylor, L., *et al.*, 2011. Talk Like an Electrician : Student Dialogue Mimicking Behavior in an Intelligent Tutoring System. *AIED 2011, LNAI 6738.* 361-368

Stenchikova, S., Stent, A., Brook, S., 2007. Measuring Adaptation Between Dialogs. *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue.* 166-173

Stern, J., Berger, J.M., 2015. *ISIS: The state of terror.* London: William Collins.

Storkel, H.L., 2001. Learning New Words: Phonotactic Probability in Language Development. *Journal of Speech, Language and Hearing Research*. **44**: 1321–1337

Storm, M., Cruickshank, P., Lister, T., 2003. *Agent Storm: A Spy Inside al-Qaeda (3rd ed.)*. London: Penguin.

Straziuso, J., 2013. Twitter Bans Al-Shabab, Somalia's Al Qaeda Extremists, For Violating Terms Of Service. *The Huffington Post.* [online]. Available at: http://www.huffingtonpost.com/2013/09/07/twitter-bans-al-shabab_n_3886236.html

Tamburrini, N., Cinnirella, M., Jansen, V.A.A., Bryden, J., 2015. Twitter users change word usage according to conversation-partner social identity. *Social Networks*. **40**:84-89.

Tchokni, S., Seaghdha, D.O., Quercia, D., 2014. Emoticons and Phrases: Status Symbols in Social Media. *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM 2014)*.

The Terrorism Act 2000. *2000 c.11.* UK

The Terrorism Act 2006. *2006 c.11.* UK

Timeline of ISIL-related events (2013), (n.d.). In *Wikipedia.* Retrieved May 25, 2016, from https://en.wikipedia.org/wiki/Timeline_of_ISIL-related_events_(2013)

Timeline of ISIL-related events (2014), (n.d.). In *Wikipedia.* Retrieved May 25, 2016, from https://en.wikipedia.org/wiki/Timeline_of_ISIL-related_events_(2014)

Timeline of ISIL-related events (2015), (n.d.). In *Wikipedia.* Retrieved May 25, 2016, from https://en.wikipedia.org/wiki/Timeline_of_ISIL-related_events_(2015)

Timeline of ISIL-related events (2016), (n.d.). In *Wikipedia.* Retrieved May 25, 2016, from https://en.wikipedia.org/wiki/Timeline_of_ISIL-related_events_(2016)

Trudgill, P., 1974. Linguistic change and diffusion: description and explanation in sociolinguistic dialect geography. *Lang. Soc*. **2**: 215-246

Twitter, (n.d.). *The Twitter User Agreement*. [online] Available at: https://g.twimg.com/policies/TheTwitterUserAgreement_1.pdf [Accessed 19 August 2016]

Twitter, 2016. *about.twitter.com/company*. [online] Available at: https://about.twitter.com/company [Accessed 19 August 2016]

UN Security Council, 2004. *Security Council resolution 1566.*

US Department of the Treasury, 2006. *U.S. Designates Al-Manar as a Specially Designated Global Terrorist Entity Television Station is Arm of Hizballah Terrorist Network*. JS-4134 [press release] 3/23/2006.

Vaux, D.L., Fidler, F., Cumming, G., 2012. Replicates and repeats—what is the difference and is it significant? A brief discussion of statistics and experimental design. *EMBO Rep.* **13**(4):291–296. doi: 10.1038/embor.2012.36

Veilleux-Lepage, Y., 2016. Paradigmatic Shifts in Jihadism in Cyberspace: The Emerging Role of Unaffiliated Sympathizers in Islamic State's Social Media Strategy. Journal of Terrorism Research. **7**(1): 36–51. doi: http://doi.org/10.15664/jtr.1183

Vesdapunt, N., Garcia-Molina, H., 2014. *Identifying Users in Social Networks with Limited Information.* 1-44

Victoroff, J., 2005. The Mind of the Terrorist: A Review and Critique of Psychological Approaches. *The Journal of Conflict Resolution.* **49**(1): 3-42.

Walton, C., 2014. *Empire of Secrets: British Intelligence, the Cold War and the Twilight of Empir*e. London: William Collins

Wang, W.S.Y., 1976. Language change. *Annals of the New York Academy of Sciences.* **280**(1):61-72

Ward, A., Litman, D., 2007. Dialog Convergence and Learning.

Webber, C., 2010. *Psychology & Crime: Key Approaches to Criminology.* London: SAGE Publications

Weimann, G., 2014. *New Terrorism and New Media*. Washington, DC: Commons Lab of the Woodrow Wilson International Center for Scholars.

Weimann, G., 2016. Terrorist Migration to the Dark Web. *Perspectives on Terrorism*. **10**(3)

Wennekers, T., Garagnani, M., Pulvermueller, F., 2006. Language models based on Hebbian cell assemblies. *Journal of Physiology-Paris.* **100**(1-3):16–30. WOS:000243319900002.

Williams, C., 2011. Twitter threatened with court over Hezbollah tweets. *The Telegraph.* 30 December [online]. Available at: http://www.telegraph.co.uk/technology/twitter/8984705/Twitter-threatened-with-court-over-Hezbollah-tweets.html [Accessed 13 February 2014]

Włodarczak, M., 2013. *Temporal Entrainment in Overlapping Speech* [PhD Thesis]. der Fakultät für Linguistik und Literaturwissenschaft: der Universität Bielefeld

Wright, S., Denney, D., Pinkerton, A., Jansen, V.A.A., Bryden, J., 2016. Resurgent Insurgents: Quantitative Research Into Jihadists Who Get Suspended but Return on

Twitter. *Journal of Terrorism Research*. **7**(2): 1–13. doi: 10.15664/jtr.1213 Available at: http://dx.doi.org/10.15664/jtr.1213

Yang, C., Srinivasan, P., 2014. Translating surveys to surveillance on social media. *Proceedings of the 2014 ACM conference on Web science - WebSci '14*. 4-12 DOI: 10.1145/2615569.2615696

Ynet., 2014. Twitter suspends Hamas's Qassam Brigades account. *Ynetnews*. 14 January [online]. Available at: http://www.ynetnews.com/articles/0,7340,L-4476517,00.html [Accessed 13 February 2014]

Zelin, A.Y., 2013. #Jihad's social media trend. *Foreign Policy*. February 5 [online]. Available at: https://foreignpolicy.com/2013/02/05/jihads-social-media-trend/ [Accessed 21 August 2016]

Zelin, A.Y., 2015. The Islamic State's model. *The Washington Post*. January 28 [online]. Available at: https://www.washingtonpost.com/blogs/monkey-cage/wp/2015/01/28/the-islamic-states-model/ [Accessed 23 August 2016]

10. Supplementary Material

10.1. Information sheet and consent form

School of Biological Sciences (Primary department)

School of Law (Secondary department)

Department of Geography (Tertiary department)

Royal Holloway, University of London, Egham, Surrey TW20 0EX, UK

Information Sheet

Thesis: Tuning in to terrorist signals
Research project: Deciding Whether Twitter Accounts Belong to the Same Person

My name is Shaun Wright and I am a PhD student at Royal Holloway University of London. I am carrying out a study about the effectiveness of algorithms to compare and match the Twitter accounts of potential Jihadists. I am supervised by:
Professor Vincent Jansen (Biology, RHUL),
Professor David Denney (Law, RHUL),
Dr John Bryden (Biology, RHUL),
Professor Peter Adey (Geography, RHUL),
Dr Alasdair Pinkerton (Geography, RHUL)

What Is The Purpose Of The Study?

Jihadists and their supporters get suspended from Twitter but often return with new accounts. To do research properly, we should match together the accounts that belong to the same people.
There are so many accounts on Twitter that it is difficult to find all of the matching accounts that belong to the same person.
Computer algorithms can match accounts together quickly but we need humans to decide whether the algorithm decision was correct.
The purpose of this study is to get you (and the other participants) to inspect the matching accounts identified by the algorithm, along with some control non-matches and decide whether or not you agree.

We also wish to investigate whether undergraduates make the same decisions in the same way as academics do when deciding whether Twitter accounts belong to the same person.

Why Have I Been Asked To Take Part?

We are looking for any undergraduates.

We cannot accept any participants who think that they might be distressed by viewing the Twitter accounts of Jihadists/ ISIS members/ terrorists/ or the supporters of terrorists.

We cannot accept any participants who support/ sympathise with/ or apologise for the actions carried out by ISIS or any other terrorist organisation.

We wish to have around 25-40 people participate in total.

What Will The Study Involve?

All participants will sit in a computer lab on campus at the same time.

You will be given the following:

(e) this information and consent sheet

(f) an instruction sheet

(g) a pile of marking sheets

(h) a stopwatch

(i) a pen (if you do not have one with you)

On the computer in front of you will be a folder containing several PDF files.

You will also be given a list of the file names that you will be opening, inspecting and making a decision about. You will be asked to inspect about 60 files (the exact number will depend on how many participants volunteer for the study).

The task involves opening each file on your list, one-by-one and inspecting the screenshots and details of Twitter accounts contained within.

You will be asked to quickly scan them and make one decision. You must decide whether the accounts within the PDF file are:

A) Full match – all of the accounts in the file belong to the same person

B) Part match – some of the accounts in the file belong to the same person, but some of them do not

C) No match – none of the accounts in the file belong to the same person

You will also be asked to note down on the marking sheet:

- Brief reasons for your decision – what features of the account / its images / its metadata / its tweets etc. convinced you to make the decision you did
- The number of accounts within the PDF file
- The number of screenshots within the PDF file
- The name of the PDF file
- Your anonymous marker ID (this is so that we can check how many different participants have inspected each PDF file).
- Time spent on this PDF file.

You will make this decision on your own, without consultation with anybody else. Seating will be arranged to prevent other participants seeing your responses. Do not worry about getting the question correct, there is no right or wrong answer.

You will have a stopwatch and be asked to spend around 1 minute inspecting and deciding about each PDF file before moving on to the next one. This will allow us to calculate the total amount of human-hours spent by all of the participants on this study.

Spending under a minute on 60 files should not take much longer than 45 minutes. With an introductory briefing at the start and a debrief at the end, the whole study should take around 1 hours. You will be free to take a break, use the bathroom, take some refreshments, or stop at any point that you would like.

There is no financial or academic benefit from participating. Free refreshments will be provided throughout.

Who Will See My Information?

Your responses in the marking sheets will be seen only by the study team listed above. Your name will only be recorded on the consent form. Your responses on the marking sheet will be anonymous. Your information and responses will be treated as confidential at all times.

You can decide not to answer some questions if you wish. You can also decide not to inspect all of the PDF files if you wish. The study will be written up and published in a scientific journal and in my thesis. Your information will not be identifiable to you when published. Any data arising from the study will only be used for the purposes of the current study. You may withdraw at any time without having to give a reason. You may also ask for your data to be withdrawn at any time without having to give a reason.

Data about how many people made each decision about each PDF file will be recorded and kept, but not who made those decisions. Data about the reasons people made decision will be recorded and kept, but not who made those decisions.
The marking sheets filled in by participants will be destroyed after the completion of the thesis and publication of the study.

Do I Have To Take Part?

You do not have to take part in this study if you don't want to. If you decide to take part you may withdraw at any time without having to give a reason. Taking part, or choosing not to take part in this study, will not affect your academic record or grade now or in the future. Withdrawing at any time will not affect your right to access the refreshments.

What Should I Do If I Would Like To Find Out More?

Please email **[redacted]**

What If There Is A Problem?

If you have a concern about any aspect of this study, you should ask to speak to the researchers who will do their best to answer your questions.

Please keep this part of the sheet yourself for reference. Please feel free to ask any questions before you complete the consent form below. The consent form will be stored separately from the anonymous information you provide for this research.

Student Counselling service?

If you find the study emotionally challenging and you don't know how to handle it, or you find that you would rather not talk to the researchers, family, friends or the department, then the Counselling Service can offer you some support.

“To see a counsellor you can either contact us by phone on **[redacted]** or drop us an email at **[redacted]** or even pop into the office at FW171 between Monday - Friday 9am - 12 pm and 1 pm - 4 pm.

In addition to an appointment with the service, we can offer a number of alternative sources of help to our current students.

Online help

We have a common problems page to give you a start at thinking about a reasonably straightforward issue such as an academic problem.

Also, if you would like some useful tips on wellbeing issues such as how to learn to relax, a helpful 10 minute podcast is available from the Mental Health Foundation on the help and information home page.
http://www.mentalhealth.org.uk/help-information/podcasts/

Someone to talk to

College subscribes to the student Nightline service (number below) which can be rung between 6pm and 8am in term-time. You will be talking to a trained student volunteer from a London University College. You can also ring the Samaritans (number below).

Nightline	020 7631 0101
email	listening@nightline.org.uk

Free calls on Skype via : www.nightline.org.uk

Health Centre	on	01784 443 131
Samaritans	on	free phone number 116 123

”

Consent form ID number…….

Deciding Whether Twitter Accounts Belong to the Same Person

You have been asked to participate in a study about whether or not accounts on Twitter that potentially belong to Jihadists belong to the same person; a study which is being carried out by Shaun Wright.

Have you (please circle yes or no):

1. Read and understood the information sheet about the study? yes no

2. Had an opportunity to ask questions? yes no

3. Got satisfactory answers to your questions? yes no

4. Understood that you're free to withdraw from the study at any time, without giving a reason and without it affecting you? yes no

5. Do you think that you might be distressed by viewing the Twitter accounts of Jihadists/ ISIS members/ terrorists/ or the supporters of terrorists? yes no

6. Do you support/ sympathise with/ or apologise for the actions carried out by ISIS or any other terrorist organisation ? yes no

Do you agree to take part in the study? yes no

Name in block letters _____

Signature_____ Date _____

NB: This consent form will be stored separately from the anonymous information you provide.

10.2. Instructions to markers – whole dataset

Please check that you have the following materials in front of you:

- A printed instruction sheet
- A printed marking sheet
- A folder (on the computer)
- A stopwatch/clock
- A pen

Instructions:

(j) Each pdf file contains information about a single Twitter user. In some cases, there will be a screenshot of the Twitter account.
In all cases, with or without screenshot, the information about the user will be summarised in a table as follows:

Set ID number:	
Handle:	*The unique @ user-name of this user (e.g. @john123456)*
Name:	*The name this user gives (e.g. John Smith)*
Biography:	*A brief, optional description the user gives themselves (e.g. I'm a Biology undergraduate at Royal Holloway. Love #football and #music.)*
Location:	*Optional self-declared location, might not actually be a place (e.g. London)*

(k) Your task is to find any groups of accounts that you believe have been created by / belong to the same person.

(l) Please spend a total of no more than 2 hours on this entire task.

(m) On the answer sheet, please use a different box for each “group of accounts”.

(n) Please note the USER ID of each account you identify.

(o) Please also briefly note down the reasons for your decision. These can be in bullet point form and you are unlikely to need more than 4 or 5 words. For example you *might* make your decision based on the profile picture, alternatively you *might* note that all users use the same unusual word in their tweets.

Brief reasons for your decision:	

(p) The following list of facts may help you in deciding whether or not a particular fact about the accounts is important to the decision about whether they match:

Names:

- *Abu* means “Father of” and is common in Arabic names
- *Umm* means “Mother of” and is common in Arabic names
- *Bint* means girl or daughter (without the negative connotations it has in English)
- *Bin* means son

Places:

- *Dar ul kufr* means the land of unbelievers, i.e. any country not governed by Muslim laws
- *Sham* also “Levant”, the region of Syria, Lebanon, Palestine, Israel, Jordan, Cyprus
- *Khorasan* an early Islamic region covering Iran, Central Asia and Afghanistan

Other:

- *Baqiya* a collection (or family) of IS supporting Twitter users around the world
- *Kafir / kuffar* means unbeliever(s) or infidel(s)
- *Ummah* the entire community of Muslims from around the world
- *Sharia* Islamic legal system derived from the Quran and the Hadith
- *Khilafah* Islamic political system/state the Prophet sought to create, implements Sharia
- *“Die in your rage”* a quote from the Quran

10.3. Marking sheet – whole dataset

Marker ID number:	

IDs in this set:	
Brief reasons for your decision:	

IDs in this set:	
Brief reasons for your decision:	

IDs in this set:	
Brief reasons for your decision:	

10.4. Instructions for markers – model marking

Please check that you have the following materials in front of you:

- A printed instruction sheet
- A printed marking sheet
- A folder (on the computer) containing 120 (or 121) pdf files
- A stopwatch
- A pen

Instructions:

(q) Each pdf file contains information about several Twitter users. In some cases, there will be a screenshot of the Twitter accounts.
In all cases, with or without screenshot, the information about the users will be summarised in a table as follows:

Set ID number:	
Handle:	*The unique @ user-name of this user (e.g. @john123456)*
Name:	*The name this user gives (e.g. John Smith)*
Biography:	*A brief, optional description the user gives themselves (e.g. I'm a Biology undergraduate at Royal Holloway. Love #football and #music.)*
Location:	*Optional self-declared location, might not actually be a place (e.g. London)*

(r) Your task is to decide whether or not you think the accounts match and belong to, the same person.

(s) There are three options on your marking sheet: Full match, Part match, No match.

Do you think this is... (please tick one)	Full match		Part match		No match	

If the pdf file only contains 2 users, then either you think that they do not match (No match), or that they do (Full match). It is not possible to have a Part match with only 2 users.

If the pdf file contains more than 2 users and you think that nobody matches anybody else, tick the (No match). If you think that everybody within the file matches everybody else within the file then tick (Full match). If however, for example, in a file containing @user1, @user2 and @user3 you think that @user1 and @user2 match with each other, but that @user3 does not match them, then tick (Part match).

(t) Please spend a total of no more than 2 hours on this entire task. Noting down start and end time.

(u) Please also briefly note down the reasons for your decision. These can be in bullet point form and you are unlikely to need more than 4 or 5 words. For example you *might* make your decision based on the profile picture, alternatively you *might* note that all users use the same unusual word in their tweets.

Brief reasons for your decision:	

(v) The following list of facts may help you in deciding whether or not a particular fact about the accounts is important to the decision about whether they match:

Names:

- *Abu* means “Father of” and is common in Arabic names
- *Umm* means “Mother of” and is common in Arabic names
- *Bint* means girl or daughter (without the negative connotations it has in English)
- *Bin* means son

Places:

- *Dar ul kufr* means the land of unbelievers, i.e. any country not governed by Muslim laws
- *Sham* also “Levant”, the region of Syria, Lebanon, Palestine, Israel, Jordan, Cyprus
- *Khorasan* an early Islamic region covering Iran, Central Asia and Afghanistan

Other:

- *Kafir / kuffar* means unbeliever(s) or infidel(s)
- *Ummah* the entire community of Muslims from around the world
- *Sharia* Islamic legal system derived from the Quran and the Hadith
- *Khilafah* Islamic political system/state the Prophet sought to create, implements Sharia
- *"Die in your rage"* a quote from the Quran

10.5. Marking sheet – model marking

Time started:	
Time finished:	
Time spent on breaks:	
TOTAL TIME TAKEN:	

Marker ID number:						
Set ID number:						
Do you think this is... (please tick one)	Full match		Part match		No match	
Brief reasons for your decision:						

Marker ID number:						
Set ID number:						
Do you think this is... (please tick one)	Full match		Part match		No match	
Brief reasons for your decision:						

Marker ID number:						
Set ID number:						
Do you think this is... (please tick one)	Full match		Part match		No match	
Brief reasons for your decision:						

END OF THESIS.