Do Socialbots Dream of Popping the Filter Bubble?
The role of socialbots in promoting deliberative democracy in social media

Tim Graham[1] and Robert Ackland[2]

[1]The University of Queensland, Brisbane, Australia

[2]Australian National University, Canberra, Australia

timothy.graham3@uq.net.au

robert.ackland@anu.edu.au

Pre-print of:

Graham, T. and R. Ackland (2016), "Do Socialbots Dream of Popping the Filter Bubble? The Role of Socialbots in Promoting Deliberative Democracy in Social Media", in M. Bakardjieva and R. W. Gehl (eds) Socialbots and Their Friends: Digital Media and the Automation of Sociality. London: Routledge.

> The electric things have their lives, too. Paltry as those lives are.
>
> —*Rick Deckard, in* Do Androids Dream of Electric Sheep?

> What counts is that we are at the beginning of something.
>
> —Deleuze (1992, p. 7)

Introduction

Philip K. Dick's seminal novel, *Do Androids Dream of Electric Sheep?* (adapted into the film *Blade Runner*), poses multiple questions about the relations between humans and non-humans. One such question concerns whether we might one day reach a future in which robotic humanoids (i.e. the titular 'androids') and humans are no longer easily distinguishable. In the age of social media, it is now evident that the question Dick initiated over half a century ago has found particular relevance in the figure of the 'socialbot'. As Gehl contends: "The last tweet you got may have been from a robot" (Gehl, 2014, p. 21). Yet ‘bots’, loosely defined as software applications involved in the automation of tasks over the internet, have existed since at least the mid-1990s. For example, web crawlers (bots that assist in the collection and indexing of web content) and ‘spambots’ (bots that send massive volumes of unsolicited ‘spam’ email) are so mundane as to appear almost invisible nowadays. Similarly, chatbots or ‘chatterbots’ (bots that engage in conversation in online spaces) have existed since the early years of the web (Mauldin, 1994), and have developed into the research area of ‘conversational agents’ (Gaglio & Lo Re, 2014, pp. 285-299). Scholars have also recently explored the role of bots in automated high-frequency trading within global financial markets, drawing to attention the world-shaking events that can emerge as a result of their complex interactions (Steiner, 2012).

Given the broader context, one might ask what is unique or interesting about socialbots. Hwang et al offer the following:

> What distinguishes these “social” bots from their historical predecessors is a focus on creating substantive relationships among human users—as opposed to financial resources—and *shaping the aggregate social behaviour and patterns of relationships* between groups of users online. (2012, p. 40, emphasis added)

In recent years a growing body of literature has explored the proliferation of socialbots in social media sites such as Twitter and Facebook. Indeed, various studies have now demonstrated that socialbots are able to infiltrate social media, remain undetected and even function ‘successfully’ as social actors (Boshmaf et al, 2011; Freitas et al, 2014). In an experiment to infiltrate Twitter using socialbots, Freitas et al (2014) found that “over the duration of the experiment, the 120 socialbots created by us received in total 4,999 follows from 1,952 distinct users, and 2,128 message-based interactions from 1,187 distinct users … a significant fraction of the socialbots acquire relatively high popularity and influence scores” (Freitas et al, 2014, p. 7). In a similar study, Hwang et al (2011) discovered that socialbots were not only able to infiltrate target sub-networks on Twitter, but also “succeed in reshaping the social graph of the 500 targets, drawing responses and interactions from users that were previously not directly connected" (Hwang et al, 2011, p. 41). Indeed, in making sense of this phenomenon, Gehl (2014) argues that socialbots are becoming enrolled in processes of *noopower* (a term drawing on Lazzarato), broadly defined as "the action before action that works to shape, modulate, and attenuate the attention and memory of subjects" (Gehl, 2014, p. 23). Emerging theoretical perspectives on socialbots suggests a subtle and complex role for social robotics in the context of social media.

The ability for socialbots to appear human-like and also *shape* social relations calls to mind the rogue *Nexus-6* androids of Dick's novel, which, in the eyes of the state, constituted a serious danger to individuals and society. Indeed, discourse in recent literature tends to construct socialbots as a kind of 'danger' or hazard to society. For example, we learn that socialbots are deployed to 'infiltrate' and 'exploit' social network sites (SNS) in order to extract or expose private information about individuals and their workplaces (see for example: Elyashar et al, 2013; Paradise et al, 2014). We are informed that 'botnets', coordinated armies of socialbots mimicking human users, are able to circumvent existing security mechanisms in order to wreak systemic havoc by spreading propaganda or misinformation (Boshmaf et al, 2011). Other studies, such as Mitter et al (2014a) have taken the dangers of socialbots into the 'meta' realm, by developing a categorisation schema to understand and counter-act the various categories of socialbot 'attacks' on SNS. There is certainly much validity to such narratives, and the negative aspects of socialbots constitute a complex, open research problem. However, there is another side to socialbots that has not attracted much scholarly inquiry, as Hwang et al argue: "While much has been made about the dark side of social robotics, several positive applications interactions of this technology are emerging" (Hwang et al, 2012, p. 40).

It is therefore evident that much research tends to highlight the dangers or risks associated with socialbots—what might be considered as the 'social bad' perspective. In this chapter we seek to evaluate the obverse of this perspective in order to explore some *beneficial* capacities of socialbots (in their capacity to 'exploit' and shape online social networks). In this way, in this chapter we tackle an idea previously raised by Hwang et al: "Swarms of bots could be used to heal broken connections between infighting social groups and bridge existing social gaps. Socialbots could be deployed to leverage peer effects to promote more civic engagement and participation in elections" (2012, p. 40). More specifically, we explore how socialbots on social media could exploit network structure to mitigate the effect of political filter bubbles and political segregation, thus promoting the Habermasian ideal of deliberative democracy – a public sphere (e.g. Habermas, 1996) where individuals can discuss matters of mutual interest and hopefully reach a common understanding or solution, or at the least can "hear the other side" (Mutz, 2006). For simplicity, we focus much of our presentation on the microblog Twitter but our ideas are applicable to any social media where people congregate to discuss and engage with political issues (e.g. web forums, fanpages and group pages in Facebook).

The remainder of this chapter is structured as follows. In the next section we define and problematise deliberative democracy in the context of the web, highlighting key theoretical perspectives and empirical research. The third section introduces and discusses the role of socialbots in promoting deliberative democracy in social media networks. In doing so, we set forth three 'principles' for socialbots that introduce key concepts and technical methods for socialbots in this role. In section four, we develop these concepts and methods further by introducing the notion of 'popperbots' and 'bridgerbots', providing a two-fold 'schematic' for programming socialbots to promote deliberative democracy in social media. Finally, we conclude with a reflection on the meaning and implications of social robotics within the entangled trajectories of politics, social media, and contemporary modes of power.

Filter bubbles and deliberative democracy on the web: network topologies, algorithmic sorting, and political homophily

> He experienced them, the others, incorporated the babble of their thoughts, heard in his own brain the noise of their many individual existences.[1]

On the web, politics unfolds through topologically variant networks, and actors both shape—and are shaped by—the hybrid socio-technological environments they co-habit. In the context of political discussion online, one might be tempted to regard the internet as an equal or neutral playing field, whereby people of all backgrounds converge to learn, debate, and participate in political discourse. This was the basis of early Utopian predictions of the impact of the web on politics (e.g. Castells, 1996): that the web would foster a new era of broad-based participation in the direction and operation of the political system. In contrast, Putnam (2000) and Sunstein (2001) predicted a loss of a common political discourse resulting from a fragmenting of the online population into narrowly focused groups of individuals who are only exposed to information that confirms their previously held opinions – later referred to as 'cyberbalkanisation' (Van Alstyne and Brynjolfsson, 2005).

These concerns about the potential impact of the web on democracy have continued into the present era of social media. In his book, ‘The Filter Bubble’, Eli Pariser argues that web users are increasingly entrapped within personal ‘filter bubbles’ that reflect back to them their already-held opinions or beliefs, and expose them to subjects they are already interested in (Pariser, 2011). The ‘filter bubble’, also referred to as the ‘echo chamber’, can be understood as emerging from two phenomena: *algorithmic sorting* (whereby external forces or 'opportunity structures' influence the types of political information and people that individuals are likely to encounter) and *individual preferences* (whereby web technologies enable individuals to efficiently select who they want to connect with and what types of information they want to be exposed to).

Algorithmic sorting occurs at both the aggregate- and individual-level. Concerns about the political implications of aggregate-level sorting first emerged in Web 1.0 research which considered the fact that the web, like many large-scale networks, has been found to exhibit a 'power law' in the distribution of inlinks (meaning a very unequal distribution, with a small number of websites enjoying many inlinks and the vast major of websites only have few or no inlinks). Hindman et al (2003) argued that power laws on the web could imply vast inequalities in the distribution of attention to different political viewpoints, since people usually find new websites either by following links (web surfing) or by using search engines such as Google, and in both cases a website is more likely to be discovered the greater the number of inlinks from other relevant sites. Aggregate-level algorithmic sorting occurs in social media in the form of “trending topics” in Twitter, for example, and forces of cumulative advantage (the 'rich get richer') can help a topic to take off. A concern is that social media companies can exert a degree of curatorial control over trending topics. An oft-cited example is the fact that, during the Occupy Wall Street movement, participants and supporters used Twitter extensively for communication and debate (garnering massive media attention), yet the #OccupyWallStreet hashtag failed to become a “trending topic” on the Twitter homepage (Gillespie, 2012).

Individual-level algorithmic sorting is undertaken by the social media providers whereby web content is ‘individualised’ based on *user demographics* (e.g. voluntarily contributed profile data or trace artefact data such as browser cookies) and/or *user activity* (e.g. what types of web content users statistically tend to be interested in). In the case of Twitter, each user has a “home timeline” that not only displays content they have elected to view, but also content that is *suggested* or *curated* by Twitter’s algorithms. As the official Twitter FAQ states: “Your home timeline displays a stream of Tweets from accounts you have chosen to follow on Twitter. New users may see *suggested content* powered by a variety of signals”. The Twitter FAQ continues: “Additionally, when we identify a Tweet, an account to follow, or other content that's popular or relevant, *we may add it to your timeline*. This means you will sometimes see Tweets from accounts you don't follow … *Our goal is to make your home timeline even more relevant and interesting*” (Twitter, 2015, emphasis added).

Dormehl (2014) argues that this ‘algorithmic culture’ has a dual nature. On the one hand, it is useful because it filters out the endless babble, or unnecessary ‘noise’, that would otherwise overwhelm users and software platforms (e.g. social media sites, search engines). But on the other hand, it is also problematic because users are not presented with “ideologically untampered” content, but rather the opposite—content that “flatter our personal mythologies be reinforcing what we already ‘know’ about particular issues”

(Dormehl, 2014, p. 47). Recent studies suggest that social media such as Facebook and Twitter are implicated in the advent of *political* filter bubbles. Whilst the extent and nature of this phenomenon is debated (Bozdag et al, 2014), the algorithmic modulation of incoming and outgoing flows of socially generated data suggests far-ranging consequences for individuals and collectives.

While individuals are to some extent guided by algorithmic sorting, the role of individual preferences in the creation of political filter bubbles is perhaps even more important. Earlier research also considered the impact for politics of the “narrowcasting” nature of the web, whereby users could use newly-invented RSS feed technology to efficiently select content from online newspapers or blogs that matched their existing political outlook. The emergence of social media has provided even more opportunity for politically-motivated social selectivity, as individuals can make conscious decisions as to who to friend in SNSs such as Facebook and who to follow, retweet or mention in Twitter. Such behaviour can lead to online networks that are highly divided along ideological or political lines, a phenomenon known as political homophily.

It is an empirical question as to whether algorithmic ‘filtering’ of content in social media (both at the scale of population-based ‘trends’ and the scale of individual user ‘timelines’) and computer-mediated social selection (friending, following, mentioning etc.) contribute to worsening already existing political divides across its network. The ‘filter bubble’ phenomenon warrants careful and serious consideration because of its possible implication in engendering creating social rifts that centre upon ideological or political lines. As Conover et al point out, “a deliberative democracy relies on a broadly informed public and a healthy ecosystem of competing ideas” (Conover et al, 2011b, p. 89).

Some Principles of Socialbots **for promoting deliberative democracy**

> "We selected her as your first subject. She may be an android. We're hoping you can tell."[2]

In this section, we identify three ‘principles’ of socialbots for promoting deliberative democracy on social media. Before enunciating our principles, it is necessary to first briefly define deliberative democracy and outline how it may be measured and quantified using network analysis.

As noted above, our definition of deliberative democracy involves the Habemasian concept of the public sphere (e.g. Habermas, 1996), an informal discursive space where where individuals and groups can reach common understanding about issues of mutual interest, thus influencing public opinion and potentially leading to political action. Our definition of deliberative democracy thus does not cover more formal deliberation that occurs at different levels within the political system (see, e.g. Dryzek 2010 for more on deliberative democracy).

A network is a set of nodes (vertices or entities) and a set of ties (edges or links) indicating connections or relations between the nodes. While there are several types of networks that can be extracted from Twitter (as noted above, the discussion focuses on Twitter for simplicity, but these ideas extend to other types of social media), we focus here on the network comprising Twitter users, where ties are created from users following each other, and retweeting, mentioning and replying to one another (we refer to this as the 'user network').

So how can we measure the extent or degree of deliberative democracy using the Twitter user network? A starting point is to construct the network of users participating in Twitter conversations on political issues, for example by only collecting tweets that feature the #auspol (Australian politics) hashtag. So the user network might consist of all Twitter users who authored at least one tweet containing #auspol, during a particular time period. A first quantitative measure of deliberative democracy is the network *modularity* score (e.g. Newman, 2006), which assesses the strength of the division of a network into “communities” (or clusters, or modules). Modularity ranges between 0 and 1, with a score closer to 0 indicating that more linking is occurring between clusters than within clusters (i.e. less balkanisation). While it is difficult to interpret a given modularity score as an absolute measure of deliberative democracy, modularity may be useful when one is comparing across networks (e.g. networks created for different political hashtags or the same hashtag, but constructed for different periods of time). So if we found that modularity score for the #auspol user network was decreasing over time then this would indicate that the Twitter conversation is becoming less clustered, thus indicating an increase in deliberative democracy.

However, underlying our use of clustering in the Twitter user network as a measure of deliberative democracy is a very strong assumption regarding the nature of interactions that are taking place in political spaces on Twitter. Specifically, our approach involves the use of large-scale unobtrusively collected digital trace data: mention, reply, retweet and follower ties. Thus, we assume that if a Twitter user creates a tie to another user (via a reply, retweet, mention or follow) then this tie either reflects a shared political outlook (political homophily) or at the very least, is indicative of a desire to engage in a considered exchange of ideas. Deliberative democracy therefore involves a qualitative dimension, that would not be accounted for in the approach we describe above. Using SNA terminology, our modularity clustering measure of deliberative democracy assumes that ties in Twitter only reflect positive affect. If members of opposing political groups started engaging in name calling or abusive behaviour on Twitter (that is, creating negative affect ties) then this would lead to a network that is less clustered, but this surely would not indicate increasing deliberative democracy.

There is a second reason why we should be careful in interpreting modularity clustering in the Twitter user network as a measure of deliberative democracy. Even if there were only positive affect ties in the network, there could still be a significant change in network modularity between two time periods without any underlying change in deliberative democracy. For example, if in the #auspol conversation on Twitter there was an increase in reciprocity (I'll retweet you because you retweeted me) or triadic closure (I follow person A and person A follows person B, therefore I'm going to follow person B too), then this could result in the #auspol network becoming more clustered (modularity score increasing) without any underlying change in political homophily.

Hence we recognise that modularity is a blunt measure of deliberative democracy, but propose it as an initial way of operationalising the principles of socialbots.

Our principles of socialbots for promoting deliberative democracy are presented in the style of Asimov's famous 'Three Laws of Robotics' (see: Asimov, 1950), however their scope and application is much less epochal or universal. The principles relate specifically to the survival and effective functioning of socialbots on social media. In the remainder of this section we expound upon these principles in more detail, before progressing to the specific roles that we envisage for socialbots (discussed in the next section).

Some Principles of Socialbots:

1. Socialbots must do no harm to human beings (measured in political and non-political terms);
2. Socialbots must protect their own existence, except where doing so would conflict with the First Principle;
3. Socialbots must make a significant improvement to deliberative democracy, obtaining *non-trivial, quantifiable effects* in the target sub-network(s), except where doing so would conflict with the First and Second Principles.

The First Principle of Socialbots

1. *Socialbots must do no harm to human beings (measured in political and non-political terms);*

Isaac Asimov ushered into the world an enduring problem in robotics—namely, that the notion of a robot causing 'harm' is very difficult to define precisely. In the context of this paper, the First Principle of Socialbots seeks to operationalise 'harm' broadly in two ways: political and non-political. We will briefly deal with both of these problems in this section and suggest several approaches to address them. Again, they are very specific to the research problem in this chapter, although we feel there may be broader applicability to socialbots vis-à-vis social media.

First, for a socialbot, what would it mean to cause *political* harm? Although this is a complex and multifaceted problem, at the most abstract level we argue that if a socialbot is positioned at a political extreme (e.g. far-right or far-left), then it is held to cause political harm and therefore contravene the First Principle. While measuring whether a socialbot is 'politically extreme' is non-trivial, we argue that this problem is not insurmountable, in light of recent developments in the literature and key concepts within social network analysis (SNA) and graph theory. We will now briefly elaborate upon two possible paths towards measuring whether, and how, socialbots could cause 'political harm' in social media networks.

First, socialbots could be programmed to endeavor to occupy a position within the target subnetwork(s) that approximates *regular equivalence* with ideologically or politically 'moderate' users. This argument centres on the graph theoretic notion of 'regular equivalence' whereby "two nodes in a social network are regularly equivalent if they fulfil the same role" (van Steen, 2010, p. 259). What we are suggesting here is that socialbots could be programmed to occupy a similar 'social role' in the network to users who are moderate in their ideological or political views. Roughly speaking, socialbots would attempt to 'blend in' by analysing the network structure of moderate users and then attempt to replicate it, aiming to maximise an approximate *regular equivalence* with such users, within the constraints of the Twitter API and computational resources of the researchers.

In order to identify which Twitter users are 'moderate' and should therefore be targeted, the methods outlined in Conover et al (2011a) or Boutyline & Willer (2014, *working paper*), appear especially suited to the task. Conover et al (2011a) find that the best way to predict political affiliation in Twitter networks is by analysing the 'community' structure of retweet networks (i.e. where nodes represent users, and links between nodes represent whether, and how many times, user i has retweeted user j, and vice versa). In their study, Conover et al (2011) manually code 1,000 randomly selected users into three political affiliation categories: 'left', 'right', or 'ambiguous'. In addition to other methods, they perform community detection on the retweet network of 23,766 users, resulting in two ‘clusters’ emerging. They classify users by political affiliation using the cluster each user is assigned to, and find that this yields a 95% accuracy when evaluated against the manually coded users. Figure 1 (below) is adapted from Conover et al (2011a, p. 197), which visualises the partisan division of the retweet network into ‘left’ (blue nodes) and ‘right’ (red nodes) clusters. We have superimposed a yellow-coloured oval where the two clusters intersect, providing a visual indication of where socialbots could look to target politically ‘moderate’ users, who bridge together the two divided clusters. The network structure of these target users would then provide a statistically calculable ‘social role’ that socialbots can emulate, by attempting to establish and maintain regular equivalence.

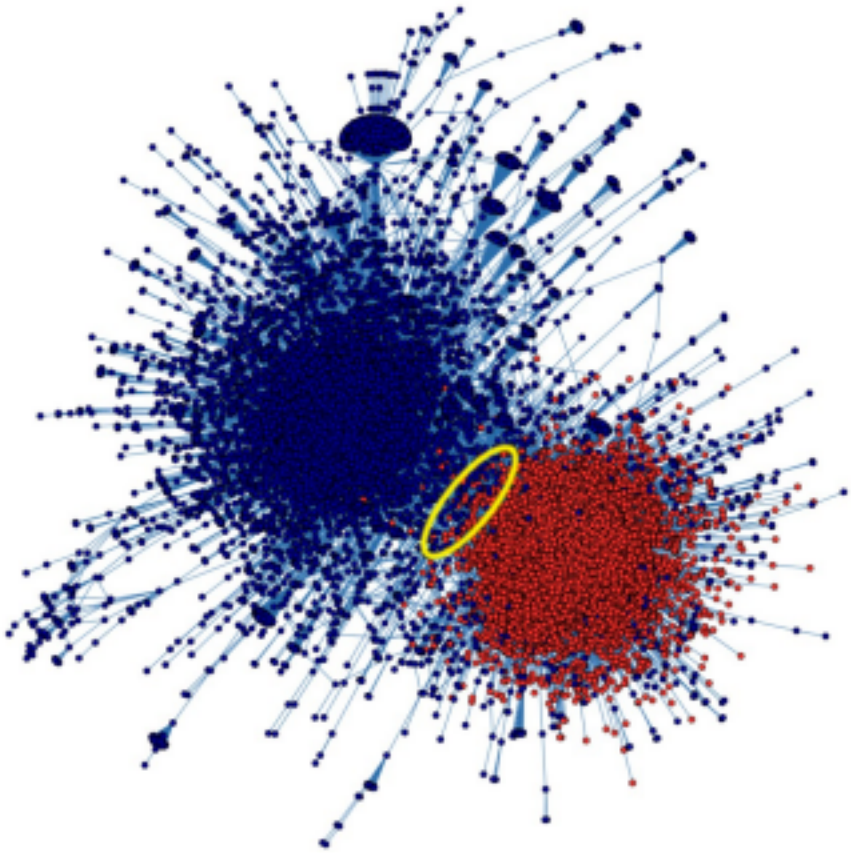

Figure 1. The political retweet network, laid out using a force directed algorithm. Adapted from *Predicting the Political Alignment of Twitter Users* (p. 197), by Conover et al, 2011, IEEE Third International Conference on Social Computing (SocialCom). Copyright 2011 by PASSAT. Adapted with permission.

Second, socialbots could be programmed to endeavor to occupy a position within the target subnetwork(s) that is *maximally neutral* in respect to quantified measures of political affiliation and/or ideological segregation. In other words, what we are suggesting is that socialbots should not find themselves in a situation where they appear to have clearly ‘taken a side’ or become, in a word, partisan. Again, it is possible to assess this, at least crudely, using the SNA methods. In order to measure whether this has occurred would necessitate programming socialbots to periodically assess the structure of their social network and their network activity. To achieve this, the techniques and methodology as developed in Conover et al (2011a), Halberstam & Knight (2014), or Golbeck (2014) would provide a suitable reference point. For example, a socialbot may calculate that its structural position within the political retweet network expresses a ‘left’ political identity, as formulated by Conover et al (2011a). In this case, the socialbot would seek to remedy this bias by retweeting from users who are calculated to have a ‘right’ political identity. Similarly, following Conover et al, socialbots may seek to ensure that their position in the network results in a classification as ‘ambiguous’, for example, by strategically ‘mentioning’ users from both clusters of the political divide (2011, p. 197) These ideas are expanded upon further in the next section, where we focus on two specific roles for socialbots for promoting deliberative democracy in social media networks.

Attention now turns to the second definition of ‘harm’ as defined in the First Principle—namely, what would it mean for a socialbot to cause *non-political* harm? Here we are concerned with a more general understanding of ‘harm’, which evokes Asimov’s enduring problematic of how to define and understand the notion of robots causing harm in a ‘social’ context. Accordingly, what we offer here is a rudimentary or preliminary path forward. We would like to focus upon one problem in particular, which has longstanding relevance to bots on the web—namely, that socialbots should never become ‘spambots’. Thus, a socialbot is said to cause harm if, through the frequency of its activity, it inconveniences other users or those managing the service. In some respects, this harkens to the ‘bad name’ or negative attention that socialbots inherit from their predecessors. Socialbot creators could take at least one of two approaches to ensure socialbots do not ‘spam’ networks and thus contravene the First Principle. The first approach could be to set fixed parameters based on evidence from the literature—for example, sending a maximum of N tweets per hour within a fixed set of times (e.g. 8am to 9pm weekdays; 1pm to 11pm weekends). Another approach would be to program socialbots to define their own parameters for 'non-spammy' update frequencies by calculating it based on other users in the network. For example, a socialbot could (periodically) query a random sample of 1000 users, calculate the average tweets per hour as a function of the *total number of status updates* and the *timestamp* of when the user was created, and then take the median value of this set of averages as a socially 'appropriate' hourly rate for sending out status updates. However, we again wish to point out that this is only one aspect of socialbots causing 'harm', for which space precludes detailed discussion in this paper. For example, determining whether the *textual content* of a given status update is 'harmful' (e.g. using offensive terms or spreading 'hate speech'). Techniques to deal with such problems may centre upon using dictionaries of terms (for offensive words) or using machine learning to build models to predict whether a tweet has a high degree of hate speech (and therefore not 'retweet' it, for example). Future research may seek to further explore such lines of inquiry.

The Second Principle of Socialbots

2. Socialbots must protect their own existence, except where doing so would conflict with the First Principle;

Perhaps the most fundamental facet of the Second Principle is that *socialbots must not be detected as non-human* (providing this does not conflict with the First Principle). However, we are not arguing that the Second Principle necessitates creating socialbots that could, for example, pass the Turing Test or instigate the kinds of existential problems presented by the *Nexus-6* androids in Philip K. Dick's novel. Far from such lofty aspirations, the Second Principle simply specifies that socialbots should present and conduct themselves in a manner that, at a minimum, ensures they survive long enough for the Third Principle to come into operation (and not contravene the First Principle). This is perhaps somewhat self-evident. Yet the scope and nature of this task is less straightforward than it might first appear, as the literature previously cited in this chapter suggests. Socialbots must not only contend with Twitter’s security mechanisms (that deploy sophisticated algorithms to find and remove fake user accounts and spambots), but also avoid ‘citizen policing’—users, organisations, or perhaps even other bots, that detect and report social robots to Twitter. And as the Third Principle serves to address, merely ‘surviving’ is only the first step for socialbots—the next problem concerns the ability to ‘thrive’. It could be argued that socialbots programmed using these Principles would simply *do nothing*, thereby satisfying the First and Second Principles. For example, a socialbot that does not send out any status updates (e.g. tweets) is arguably following an optimal strategy to avoid detection and do no harm. However, the Third Principle (below) ensures that this situation cannot occur, or, in the case that it does, there is logical reason for such inaction.

Furthermore, to achieve the Second Principle (and arguably the Third Principle), socialbots must present and conduct themselves in a manner that makes them appear sufficiently ‘human’ to, for example, attract new followers and retweets (again, without contravening the First Principle). Although previous studies have achieved success with the ‘detection avoidance’ problem, the problem of how to exploit social networks for optimal effect proves trickier. For example, some studies suggest that female socialbots with ‘attractive’ or ‘good-looking’ profile photos are more successful for social engineering on SNS (Boshmaf et al., 2011). Others find that the ‘gender’ of socialbots has no correlation with success or popularity (Freitas et al., 2014). Still others, such as Wald et al (2013), take a different tack by looking at which types of human users socialbots should target for interaction. Wald et al found that the highest predictors of whether a user is likely to interact with socialbots comes down to how popular or influential a user is (i.e. their ‘Klout’ score and number of friends), and the amount of sexual language and terminology they tend to use (Wald et al, 2013, p. 10). The implication is that users who are more likely to interact with socialbots (e.g. retweeting or ‘liking’ their tweets) are those that are well-connected or have more followers, and those that use a greater amount of sexual language and terminology. Clearly, in terms of SNA methods, ensuring that socialbots function effectively in social media networks involves both ‘art’ and ‘science’. At the same time, it reinforces the importance of the First and Second Principles as one way to approach socialbot ethics.

The Third Principle of Socialbots

3. Socialbots must make a significant improvement to deliberative democracy, obtaining *non-trivial, quantifiable effects* in the target sub-network(s), except where doing so would conflict with the First and Second Principles.

At an abstract level, the Third Principle seeks to ensure that socialbots are actually achieving *something* (providing it does not contravene the First and Second Principles). In this way, socialbot activity must be *quantifiable* (accounted for statistically) and must also be *non-trivial* (having a magnitude of effect that is not negligible). It is therefore evident that analysis of the impact or effects of socialbots must pay attention to network structure and network dynamics over time. Any studies that investigate whether socialbots could, for example, heal social rifts, promote deliberative democracy, bridge segregated subnetworks, or 'pop' filter bubbles, must be able to formalise socialbot activity as a *concrete, statistically calculable phenomenon*. A growing body of literature demonstrates that the methods and formalisms of SNA provide such tools. More specifically, SNA methods to quantify and analyse political segregation and ideological clustering on Twitter have emerged in recent years (see: Conover et al, 2011a; Halberstam & Knight, 2014; Golbeck, 2014). In particular, Mitter et al (2014b) provide a detailed methodology for assessing the impact of socialbots 'attacks' on Twitter in terms of shaping or influencing the social graph of a subset of users. Any combination of these methods would be suited to advancing the Third Principle of socialbots, and such methods are expanded upon later in this section. Furthermore, to achieve the Third Principle, socialbots must present and conduct themselves in a manner that makes them appear sufficiently 'human' to, for example, attract new followers and retweets. This is consistent with the Second Principle, and again, must not be in contravention of the First Principle.

We can further operationalise the Third Principle by making the following argument: the presence of socialbots in target sub-networks should, over time, correlate with a *decreased modularity score* (thus implying decreased political homophily in the target sub-network, although noting our caveat about equating changes in modularity with changes in homophily). This brings us deeper into the realm of socialbot ethics and further reveals the *raison d'être* of socialbots in promoting deliberative democracy. In this way, we can begin to explicate the 'life goals' or *telos* of socialbots in the context of this paper—broadly speaking, to build bridges between separate, ideologically homogeneous subnetworks; to expose tightly knit clusters of users to alternative viewpoints; or to bring about measurable shifts towards deliberative democracy in online discourse. In this way, the Third Principle draws stark attention to the *normative political rationalities* that socialbots in this role embody—which could be conceived as a kind of social robotic 'hacktivism'. As Howard (2003) writes, hacktivism is understood broadly as using the tools and strategies of hackers for political ends: "hacktivists believe that they have a responsibility to expose abuses of power and to *redistribute informational resources*" (Howard, 2003, p. 216, emphasis added). Positioning socialbots as ersatz 'hacktivists' facilitates a rethinking of their agential capacities—in this case, to propagate deliberative democracy on social media via the strategic exploitation of network structure.

***Popperbots and bridgerbots*: a schematic for programming hacktivist socialbots on Twitter**

> In .45 of a second an android equipped with such a brain structure could assume any one of fourteen basic reaction-postures.[3]

Programming bots to perform social roles in social media environments represents a moving target. Over time, the tasks to be performed by socialbots become suboptimal or even impossible in environments whereby the entities involved—users, protocols, algorithms, data, hardware specifications, and so forth—are constantly in flux. However, the aim in this section is not to provide a comprehensive or codified tutorial for programming socialbots, but rather to set forth a general 'schematic' for how socialbots might be programmed to promote or 'propagate' deliberative democracy on Twitter. We wish to focus on issues of methodology and the conceptual, network-oriented space in which such methods would be applied, which are broadly located at the intersection of politics, the dynamics of social media networks, and social robotics. We want to examine some possibilities and sketch out possible approaches moving forward. The over-arching question asks whether it is possible to program socialbots to mitigate or break down political filter bubbles and ideological segregation on Twitter, hence promoting deliberative democracy in online discourse. In this section we provide a possible answer to this question by elucidating two distinct roles for socialbots.

1. 'Popperbots'

We conceive the 'popperbot' as a type of socialbot tasked with the role of 'infiltrating' subnetworks of Twitter users that exhibit *high or extreme levels of homophily*. Once the popperbot has established itself in the subnetwork, it would then begin to 'inject' information reflecting more moderate or even contrasting ideological standpoints. As the name suggests, the idea is that this type of socialbot will reduce, or in a sense 'pop' the ideological bubble that users within a given subnetwork are situated within, by exposing these users to alternate points of view that appear to come from a member of their own cohort. The *telos* of the popperbot is to produce measurable increases in *heterophily* in the subnetworks in which they have infiltrated. Similarly, as argued previously in relation to the Third Principle, popperbots could attempt to decrease 'balkanisation' by striving to reduce the modularity score of their target subnetwork. For example, a popperbot could be programmed to occasionally (say, with probability P) *retweet* or *reply to* users from a different subnetwork(s) that represent alternate positions on some issue. *Figure 2* (below) shows a popperbot infiltrating a homophilous subnetwork of Twitter users who are calculated to be 'right' (i.e. conservative) in their political orientation, which, as mentioned in the previous section, could be derived using the methods outlined in Conover et al (2011a).

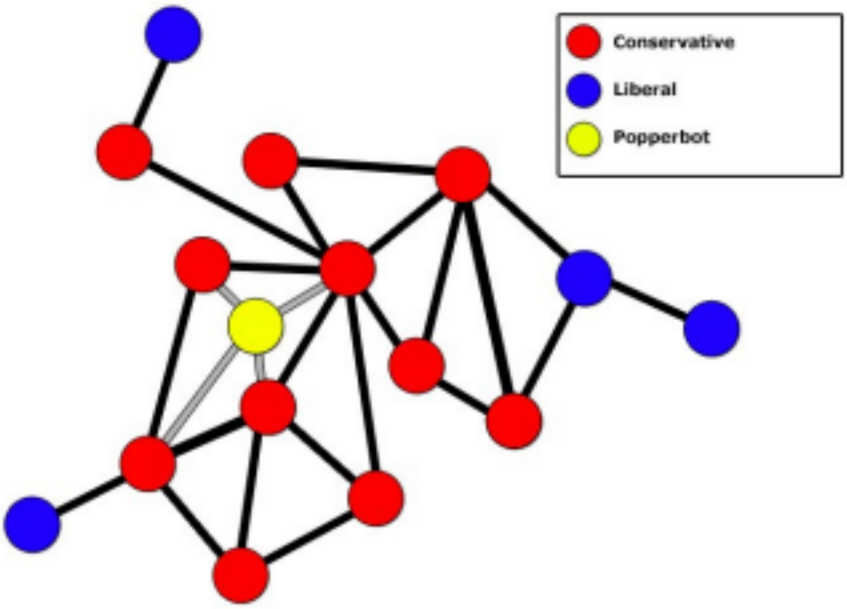

Figure 2: a ‘popperbot’ infiltrating a homophilous subnetwork (of politically conservative Twitter users)

If it is not yet apparent, a fundamental problem for the development of popperbots is that these socialbots must, by definition, violate the First Principle in order to do their job. In this way, a popperbot would have to act in an ‘extreme’ manner (as discussed previously) in order to infiltrate a homophilous subnetwork, *even if its ultimate goal was to ‘pop’ the political bubble* in that subnetwork. For example, a popperbot could (1) detect an extremely homophilous subnetwork of individuals and then (2) attempt to infiltrate the network by adapting its profile and ‘social’ activity to correspond with the target subnetwork. This popperbot could be (3) programmed to ‘defect’ after some time duration T or acquiring a pre-defined number of followers or friends N. Defection, of course, would occur in the form of injecting more moderate or perhaps even contrasting information into the subnetwork, as discussed previously. However, a popperbot would never infiltrate a network by pretending to be, for example, a radical Communist because doing so would definitely (and as we have previously argued, *quantifiably*) violate the First Principle. Thus, Asimov’s enduring problem remains and we inherit another complex, or perhaps ‘wicked’, problem to address. Yet, despite these obstacles, we argue that there are possibilities for moving forward, which could be programmed into popperbots. For example, future research could explore lines of inquiry centering on time-limited infiltration and defection routines, which could allow socialbots to violate Principles within certain parameters or ‘thresholds’ of violation, although space precludes further discussion in this chapter.

2. ‘Bridgerbots’

‘Bridgerbots’ are conceptualised as socialbots tasked with the role of re-routing or ‘bridging’ informational flows between otherwise *ideologically segregated* sub-networks. They could perform actions such as tweeting/retweeting and following users from both ‘sides’ of a given political or ideological debate. Bridgerbots would seek to expose users from one homophilous subnetwork to politically diverse types and flows of information from one or more other homophilous subnetworks. In this way, the network role of bridgerbots might be thought about in a variety of ways. One possibility is in terms of what Mark Granovetter described as *weak ties*. Weak ties are understood as connections between different tightly knit groups that are vital to information dissemination and therefore social opportunities. As Granovetter wrote, “It is remarkable that people receive crucial information from individuals whose very existence they have forgotten” (Granovetter, 1973, p. 1372).

Bridgerbots could be programmed to endeavor to occupy a position within the target subnetwork(s) that maximizes their own *betweenness centrality* score. Betweenness centrality, or simply ‘betweenness’, is a key concept in SNA and graph theory more broadly. In a formal sense, the “*betweenness* sigma(M) of a vertex *m* is the total number of shortest paths between all possible pairs of vertices that pass through this vertex” (Dorogovtsev & Mendes, 2003, p. 18, emphasis original)[1]. We can think about betweenness in terms of how important a node (a.k.a. vertex) is in providing a path that connects isolated nodes or isolated clusters of nodes. Thus, informally, nodes with high betweenness could be loosely conceived as ‘brokers’ or ‘exchange terminals’ of information between densely connected (or ‘homophilous’), but otherwise poorly connected clusters of individuals. The application of this concept to bridgerbots is straightforward—they would seek to act as ‘bridges’ between

[1] In graph theory, a ‘path’ is an unbroken sequence of connections between two or more vertices.

politically segregated clusters of users. Figure 3 (below) visualises this idea by representing it within a graph.

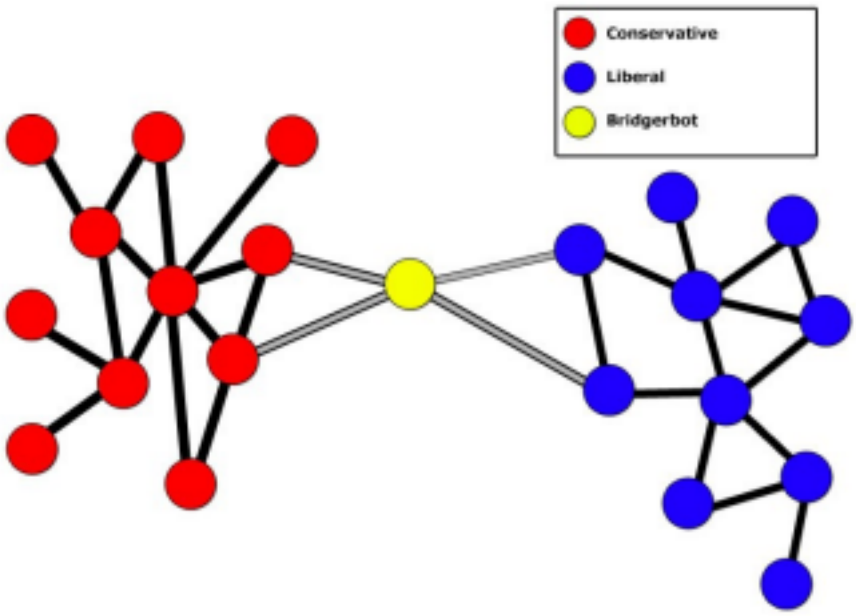

Figure 3: a ‘bridgerbot’ connecting two ‘segregated’ subnetworks (of politically conservative Twitter users)

Combining several arguments presented thus far, the role and effects of bridgerbots in respect to the target sub-network(s) could be tied to their success in *increasing* their own betweenness score or *decreasing* the modularity score of the subnetwork that they target. That is to say, bridgerbots with *high betweenness scores* are ‘bridging’ political rifts more effectively than those with a low score, and bridgerbots who successfully *decrease the modularity score* of a target subnetwork(s) are successfully ‘bridging’ political divides. For example, we could imagine that a Twitter user receives a notification that a new user has ‘followed’ them—and they might even return the gesture by ‘following back’. Unbeknownst to the user, they are now following a bridgerbot who has ‘targeted’ them because their social network is extremely homophilous. Later, having possibly completely forgotten about this new social connection (to the bridgerbot), the user might notice a tweet in their news feed that reflects a more moderate, or perhaps even competing, position to their already held beliefs on some issue. In this way, the bridgerbot has acted as a weak link, bridging heterogeneous flows of information (e.g. two or more different sides of a debate) to the actors involved in such communication networks. Over time, if the bridgerbot's betweenness score increases, or the network structure becomes less modular, then the bridgerbot can be regarded as doing a 'good job'.

Conclusion

> "Your position, Mr. Deckard, is extremely bad morally. Ours isn't."[4]

In this chapter we have introduced and examined a role for socialbots that positions them not as dangers or annoyances, but rather as *socially beneficent* actors, capable perhaps of 'building a better world'. We have focussed upon a *normative* role for socialbots in creating and propagating deliberative democracy on Twitter. This, in one sense, can be thought about as socialbots 'popping the political filter bubble'—e.g. building bridges between separate, ideologically homogeneous subnetworks, exposing tightly knit clusters of users to alternative viewpoints, or bringing about measurable shifts towards deliberative democracy in online discourse. Yet, if socialbots 'dream' of popping filter bubbles, perhaps we can perceive within their dreams the spectre of our own political rationalities and ethical assumptions. As Paul Henman writes, "new and emerging technologies will continue to initiate old questions in new circumstances of what these technologies mean" (Henman, 2013, p. 300). It is clear that socialbots of the kind we conceive in this chapter might also dream in *other* ways, ways that might otherwise seem unethical or politically abhorrent.

In developing this chapter, we consciously adopted a normative position for socialbots in relation to a particular social issue (deliberative democracy). In doing so, this provided a space in which to demonstrate and examine how socialbots might be used to 'exploit' network structure in order to achieve 'social good'. Yet what is defined as ethical, politically rational, socially beneficent, etc., arguably depends upon one's point of view. Hence, we can see how socialbots could be deployed to achieve different or even *opposite* outcomes for deliberative democracy, by simply adapting or perhaps 'inverting' various aspects of the ideas and methods established in this chapter. As Hwang et al would have it: “The same bots that can be used to surgically bring together communities of users can also be used to shatter those social ties. The same socialbot algorithms that might improve the quality and fidelity of information circulated in social networks can be used to spread misinformation” (Hwang et al, 2012, p. 40).

A particularly noteworthy focus is governments seeking to monitor and sway political discourse online. For example, the *50 Cent Party* are “Party-paid internet commentators and opinion guiders” (Sullivan, 2012) hired by the Chinese government and other parties to attempt to steer online opinion and conversation towards particular directions. Yet one can easily imagine the *50 Cent Party* deploying socialbots alongside, or even in lieu of, human commentators and opinion guiders. Indeed, as Gehl writes, government agencies such as the U.S. Air Force have already begun to contract out software development companies to “gather intelligence, build consensus, and influence opinions twenty-four hours a day via a network of socialbots” (Gehl, 2014, p. 39). In this way, current concerns regarding the uses and abuses of socialbots within a political context are not unfounded. However, it is also clear that we are only in the very early stages of this phenomenon. Hence, what we are now witnessing is an increasing sophistication of socialbot technologies and a diversification of their roles and relations of power in hybrid techno-social environments (see Gehl, 2014). Gilles Deleuze once wrote: “What counts is that we are at the beginning of something” (Deleuze, 1992, p. 7). The question is how it will unfold.

Acknowledgements

We gratefully acknowledge the helpful feedback provided by Paul Henman on an earlier draft of this chapter.

Notes

The quotations commencing each section are taken from Philip K Dick's novel *Do Androids Dream of Electric Sheep?* The corresponding page numbers are:

[1] Page 11;

[2] Page 22;

[3] Page 25 (spoken by the character *Eldon Rosen*); and

[4] Page 14.

References

Asimov, I. (1950). I, Robot. New York: Doubleday & Company.

Boshmaf, Y., Muslukhov, I., Beznosov, K., & Ripeanu, M. (2011). The socialbot network: When bots socialize for fame and money. Paper presented at the 93-102.

Boutyline, A. & Willer, R. (2014, working paper). 'The Social Structure of Political Echo Chambers: Ideology and Political Homophily in Online Communication Networks'. Retrieved 28 March, 2015, from https://www.ocf.berkeley.edu/~andrei/downloads/echo.pdf.

Bozdag, E., Gao, Q., Houben, G., & Warnier, M. (2014). Does offline political segregation affect the filter bubble? An empirical analysis of information diversity for Dutch and Turkish twitter users. Computers in Human Behavior, 41, 405-415.

Butts, C. T. (2007). Social network analysis with sna. Journal of Statistical Software, 24(6).

Castells, M. (1996). The rise of the network society. The information age: Economy, society and culture Vol. I. Blackwell, London.

Conover, M.D., Goncalves, B., Ratkiewicz, J., Flammini, A., Menczer, F. (2011a). 'Predicting the Political Alignment of Twitter Users', Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), pp. 192-199.

Conover, M.D., Goncalves, B., Ratkiewicz, J., Flammini, A., Menczer, F. (2011a). ‘Political polarization on twitter’, Proceedings of the 5th International Conference on Weblogs and Social Media.

Deleuze, G. (1992). ‘Postscript on the Societies of Control’, October, Vol. 59, (Winter 1992), MIT Press, Cambridge, MA, pp. 3-7.

Dick, P. K. (1999)[1968]. Do androids dream of electric sheep?. London: Millennium.

Dormehl, L. (2014). The Formula: How algorithms solve all our problems… and create more. Penguin Publishing Group.

Dorogovt͡sev, S. N., & Mendes, J. F. F. (2003). ‘Evolution of networks: From biological nets to the internet and WWW’. Oxford; New York: Oxford University Press.

Dryzek, J. S. (2010). *Foundations and Frontiers of Deliberative Governance*. Oxford: Oxford University Press.

Elyashar, A., Fire, M., Kagan, D., & Elovici, Y. (2013). Homing socialbots: Intrusion on a specific organization's employee using socialbots. Paper presented at the 1358-1365.

Freitas, C. A., Benevenuto, F., Ghosh, S., & Veloso, A. (2014). Reverse engineering socialbot infiltration strategies in twitter. arXiv preprint arXiv, 1405.4927.

Gehl, R. W. (2014). Reverse engineering social media: Software, culture, and political economy in new media capitalism. Philadelphia, Pennsylvania: Temple University Press.

Gillespie, T. (2012). Can an algorithm be wrong? Retrieved 22 March, 2015, from http://limn.it/can-an-algorithm-be-wrong/

Golbeck, J., & Hansen, D. (2014). A method for computing political preference among twitter followers. Social Networks, 36, 177.

Granovetter, M. (1973). The strength of weak ties. *American Journal of Sociology*, 78(6), pp. 1360-1380.

Habermas, J. (1996). Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy. MA: MIT Press.

Henman, P. (2013) Government and the internet: evolving technologies, enduring research themes, pp. 283 – 306 in Dutton, W. H. (ed.) The Oxford Handbook of Internet Studies. Oxford University Press.

Hindman, M., Tsioutsiouliklis, K., and Johnson, J. (2003). ‘Googlearchy’: How a few heavily-linked sites dominate politics on the Web. Mimeograph, Princeton University, 2003.

Howard, P. (2003). Hacktivism. *In* Jones, S. (Ed.), *Encyclopedia of new media.* (pp. 216-217). Thousand Oaks, CA: SAGE Publications, Inc.

Hwang, T. Pearce, I., and Nanis, M. (2012). Socialbots: voices from the fronts. Interactions 19, 2 (March 2012), 38-45. Retrieved 5 March, 2015, from http://doi.acm.org/10.1145/2090150.2090161.

Mauldin, M. L. (1994). Chatterbots, TinyMUDs and the Turing Test: Entering the Loebner Prize Competition. Proc. AAAI-94. URL accessible via: http://aaaipress.org/Papers/AAAI/1994/AAAI94-003.pdf

Mitter, S., Wagner, C., & Strohmaier, M. (2014a). A categorization scheme for socialbot attacks in online social networks.

Mitter, S., Wagner, C., & Strohmaier, M. (2014b). Understanding the impact of socialbot attacks in online social networks.

Mutz, D. (2006). Hearing the Other Side: Deliberative Versus Participatory Democracy. New York: Cambridge University Press.

Newman, M. E. J. (2006). Modularity and community structure in networks. Proceedings of the National Academy of Sciences of the United States of America, 103(23), 8577-8582.

Paradise, A., Puzis, R., & Shabtai, A. (2014). Anti-reconnaissance tools: Detecting targeted socialbots. IEEE Internet Computing, 18(5), 11-19.

Pariser, E. (2011). The filter bubble: What the internet is hiding from you. New York: Penguin Press.

Putnam, R. D. (2000). Bowling Alone. Simon & Schuster, New York.

Steiner, C. (2012). Automate this: How algorithms came to rule our world. New York: Portfolio/Penguin.

Sullivan, J. (2012). A tale of two microblogs in china. Media, Culture & Society, 34(6), 773-783.

Sunstein, C. (2001). Republic.com. Princeton University Press, Princeton.

Twitter, Inc. (2015). *What’s a Twitter timeline?*. Retrieved 1 April, 2015, from https://support.twitter.com/articles/164083-what-s-a-twitter-timeline

Van Alstyne, M. and Brynjolfsson, E. (2005). Global village or cyber-balkans? Modeling and measuring the integration of electronic communities. Management Science, 51(6):851–868.

Wald, R., Khoshgoftaar, T. M., Napolitano, A., & Sumner, C. (2013). Predicting susceptibility to social bots on twitter. Paper presented at the 6-13.