Contents lists available at ScienceDirect

Online Social Networks and Media

journal homepage: www.elsevier.com/locate/osnem

The strength of weak bots

Marijn A. Keijzer [a,*], Michael Mäs [a,b]

[a] ICS/Department of Sociology, University of Groningen, The Netherlands
[b] Institute of Technology Futures, Department of Sociology, Karlsruhe Institute of Technology, Germany

ARTICLE INFO

Keywords:
Social bots
Misinformation
Fake news
Social influence
Agent-based model

ABSTRACT

Some fear that social bots, automated accounts on online social networks, propagate falsehoods that can harm public opinion formation and democratic decision-making. Empirical research, however, resulted in puzzling findings. On the one hand, the content emitted by bots tends to spread very quickly in the networks. On the other hand, it turned out that bots' ability to contact human users tends to be very limited. Here we analyze an agent-based model of social influence in networks explaining this inconsistency. We show that bots may be successful in spreading falsehoods not despite their limited direct impact on human users, but because of this limitation. Our model suggests that bots with limited direct impact on humans may be more and not less effective in spreading their views in the social network, because their direct contacts keep exerting influence on users that the bot does not reach directly. Highly active and well-connected bots, in contrast, may have a strong impact on their direct contacts, but these contacts grow too dissimilar from their network neighbors to further spread the bot's content. To demonstrate this effect, we included bots in Axelrod's seminal model of the dissemination of cultures and conducted simulation experiments demonstrating the strength of weak bots. A series of sensitivity analyses show that the finding is robust, in particular when the model is tailored to the context of online social networks. We discuss implications for future empirical research and developers of approaches to detect bots and misinformation.

1. Introduction

Since the 2016 US presidential election, there is growing attention for an ancient political weapon: misinformation. Pundits and scholars fear that various actors attempt to manipulate news media and, in particular, online social media platforms. While multidisciplinary research has generated much insight into attackers' approaches to influence news media, users' perception of manipulated content, and its diffusion in online social networks, there is also a growing body of seemingly contradictory findings. With a taste for irony, Ruths [1] recently pointed out that *"the field of research on misinformation has come to resemble the very thing it studies"*.

Social bots – automated social-media accounts programmed to influence users' opinions and public discussions – have been identified as a key approach to spreading misinformation in networks. Estimates show that, in the months leading up to the 2016 US presidential election, over 400,000 bots were active in political discussions on Twitter, accounting for a fifth of the total number of tweets in this period [2]. A number of these bots focused on spreading misinformation–statements or articles that contain factually incorrect information [1,3]. Platforms that produce misinformation often use social bot accounts to amplify the early spreading of content [4]. The US Senate Intelligence Committee concluded that the Russian government deployed social bots to spread false information and falsehoods to influence the election outcome [5].

Responding to prominent calls for empirical research into the impact of bots on opinion formation and public debate [e.g. 6,7], empirical researchers found two seemingly inconsistent empirical patterns [e.g. 3,4,8,9]. On the one hand, it turned out that bots tend to be well connected to each other, but only to a few human users [e.g. 4], and that bot's direct influence on those human users seems limited [9, 10]. On the other hand, bots' messages tend to propagate through social media platforms quickly and easily [e.g. 3], reaching, and potentially influencing a large portion of social media users. How is it possible that bots are only weakly embedded in the social network, yet they have a disproportionately large impact on opinion dynamics in the network?

Explanations solving this puzzle have been sought in characteristics of social-network users, properties of misinformation, and characteristics of the communication context. For instance, it was found that misinformation is accepted more readily by individuals scoring low on analytic thinking tests, suggesting that fake news spread very fast in parts of the network where users tend to credulously accept new information [11]. Likewise, it has been argued that fake news spreads quickly in a network once it has entered because it tends to be negative,

[*] Corresponding author.
E-mail address: m.a.keijzer@rug.nl (M.A. Keijzer).

https://doi.org/10.1016/j.osnem.2020.100106
Received 30 April 2020; Received in revised form 31 August 2020; Accepted 3 November 2020
Available online 23 December 2020

shocking, and emotional. This motivates users to engage with fake news and share it with other users [12]. Some may even buy into an unbelievable story because it fits their partisan preoccupation [13], or because individuals communicate faster, more sloppy, and less considerate on online social networks than in other communication contexts [14].

While these individual-level explanations certainly contribute an important part to solving the puzzle why bot-emitted fake news seems to have a significant impact on public discourse despite bots' low network embeddedness, they neglect the complexity arising from the interaction of actors on the local-level of social networks [15]. In a social network, the impact of a node on its neighbors may be small, but each neighbor is exerting influence on another set of nodes, potentially sparking chain reactions that can spiral into large effects on the network as a whole. Counter-intuitively, modeling work on opinion dynamics in social networks even suggests that actors exerting relatively weak influence on their direct network neighbors can actually have a bigger impact on the distribution of opinions in the overall network [16,17]. This suggests that bots may not be influential despite their low embeddedness but because of their limited influence on their direct network neighbors.

The social mechanism generating this counter-intuitive effect is straightforward. Consider a bot emitting content to a group of users who consider the bot a reliable source. Influenced by the bot's content, these users will adjust their beliefs, growing more similar to the views advocated by the bot. As a consequence, these users distance themselves from those who are not directly exposed to the bot's extreme beliefs. These increased opinion differences, in turn, will decrease the influence that friends of the bot can exert on their friends. If the bot continues to "pull" its connections towards its position, their friends may refuse to be influenced any longer. Thus, while the bot had a strong influence on its direct contacts, it failed to exert indirect influence on its friends' friends, their friends, and so on. Consider, in contrast, a bot exerting only weak influence on its direct contacts and managing to pull their opinions only slowing into its direction. These users will remain able to influence their friends, pulling them also slowly but gradually into the direction of the bot's opinion. This process may take longer, but eventually the bot will not only have manipulated the beliefs of its direct network neighbors but also to a larger extent those of its indirect contacts. In other words, limited influence on directly connected users may foster the influence on indirectly connected network users.

In this study, we demonstrate the counter-intuitive effectiveness of seemingly ineffective bots in a series of computer simulations with an agent-based model. In an agent-based model, researchers build an artificial world and make assumptions about the behavior of individual actors (called "agents") and how they interact with their environment [18]. In particular, agent-based models make it possible to study the complexity arising when agents respond to each other, and chains of reaction lead to complicated phenomena that would have remained hidden without a formal analysis of the model. In an agent-based model of an online social network, for instance, one specifies how users and bots create and share content, and how they adjust their opinions after exposure to content they receive. Next, analytical or computational methods are used to study the dynamics that these assumptions generate. Here, we study a simple model of an online social network of human users and a bot, building on Axelrod's famous model of cultural dissemination [19]. This model is particularly well suited for the study of social bots and their effect in online social networks, as it can capture how content emitted by an agent can diffuse through a network.

Our analyses also revealed a surprising bot-effect. We found that highly active bots do not only fail to influence their indirect contacts but also influence fewer of their direct network neighbors than bots with a low rate of activity. We argue that this effect emerges because bot's direct contacts may adopt bot content but likely drop it when their friends fail to reinforce it. As strong bot's friends fail to convince their friends of the bot messages, this affirmation is missing.

The remainder of this paper is organized as follows. In the next section (Section 2), we reflect on the current state of the literature on the automated spreading of misinformation. Subsequently (in Section 3), we describe a formal model of social influence and the dissemination of beliefs in networks [19]. We present the results from a series of simulation experiments (in Section 4) and reflect on the main findings and implications for science and policy (in Section 5).

This paper yields two main take-away messages for engineers of social media platforms, policymakers, and scientists concerned with automated spreading of misinformation: (1) The number of bots trying to influence public debate and the number of messages they emit may not be as important as it seems. Bots that appear to have only limited impact on directly connected users can have a stronger impact on the collective opinion dynamics as they exert stronger indirect influence on the friends of their friends. (2) Detecting influential bots programmed to manipulate opinion formation and online debate may even be harder than researchers expect. State-of-the-art bot detection algorithms claim to achieve impressive detection rates around 90 percent [e.g.2,20]. However, the most ingeniously engineered bots are likely the ones who are harder to detect, and as those may have a powerful impact on the spreading of falsehood, attempts to detect these accounts or fact check their content could come in vain.

2. Background

Whether and how social bots are involved in the spreading of misinformation in online social networks has received plenty of scholarly attention in the past years [1]. Researchers consistently observe that social bots, automated social media accounts developed to manipulate processes of opinion formation and online debate, are omnipresent [1, 6].

Some empirical work suggests that bots form a threat to opinion formation in online social networks. An analysis of 14 million messages related to 400 thousand news articles on Twitter found that a disproportionate amount of tweets promoting low credibility sources came from accounts that were likely automated [4]. While bots mostly tweet amongst themselves, it has also been observed that they play a crucial role in the early amplification of information spreading [4]. The information they emit appears to be very appealing to users. A randomized field experiment showed a considerably larger reach for misinformation on online social media [3]. Fake news proves particularly potent in the fast media consumption environments that are social media platforms, where users make limited cognitive capacity available when evaluating the validity of information they encounter [14]. Users may be likely to pass on content that is negative, shocking, and emotional [12], or fits their currently held beliefs [13].

Other researchers concluded that the influence of bots is largely overstated. An analysis of registered voters on Twitter showed that a mere 0.1% accounted for 80% of sharing from fake news sources, and 1% of individuals included in the data accounted for 80% of fake news exposures [8]. Recently, another team of researchers was able to assess opinion influence directly and found that the Russian Intelligence Agency failed to exert direct influence in their sample [9].[1]

While the finding that bots can emit their content only to a small number of human users is an important empirical observation, it may be misleading to conclude that bots have only limited influence on opinion formation in the social network. Online social networks are complex systems with millions of users emitting, evaluating, adjusting, and responding to vast amounts of content [e.g.15]. In such systems, even seemingly small events can spark chain reactions that have a huge impact on the system as a whole [e.g.21]. Predicting such chains of reaction and their outcomes is highly challenging. For instance, it turns

[1] We note that this analysis includes only 12 users linked to IRA accounts, and took place a year after the election these accounts aimed to disturb.

out that no one influencer or message characteristic can be used to predict the reach of a message [22]. In his call to action, Ruths [1] argued that this re-sharing of ideas is a *"key blind spot"* in the field, and claims that *"there is a serious need for a better understanding of how fake-news stories transform into rumors and to what extent these rumors can amplify beliefs and infiltrate other communities"* [1].

Ruth's proposal seems to be that beliefs can travel through a network beyond the direct influence of the account, post, or tweet that seeded it. Ultimately, the warning for *rumors* expresses a fear of a rapid rise of uncontrollable interpersonal influence between individuals that did not even see the original post. This idea resonates with earlier work on the bounded-confidence model which showed that radicals and opinion leaders can be successful at persuading a full population through indirect influence pathways [16,17]. Subtle and gentle persuasion through limited and well-timed interaction events persuade smaller fractions of a population at a time, allowing for more influential interactions between the bot's direct and indirect contacts. Over time, indirect influence then allows the bot to attract larger shares of the total population.

Intuitively, one would expect that more connected and more active bots are also more successful at propagating their beliefs. Counter this intuition, we argue that the opposite may be true, in that bots communicating infrequently and only to a few human users may actually be more successful in spreading their beliefs. Here, we put these competing *intuitions* to the test, analyzing their logical validity with an agent-based model. In particular, we investigate the conjectures that (1) weakly connected bots are not necessarily less effective at propagating falsehoods and that (2) social bots are more effective when they are emitting content infrequently. We refer to these conjectures as the *strength-of-weak-bots* effects.

The name of the strength-of-weak-bots effects reminds one of Granovetter's famous strength-of-weak-ties argument [23]. He argued that humans often profit more from weak network-ties rather than from their very close, strong social relationships, because weak ties connect them to more diverse individuals and, thus, provide access to information that strong ties fail to provide. One the one hand, the strength-of-weak-bots effect resembles Granovetter's notion in that a seemingly weak aspect turns out to actually be a strength. On the other hand, unlike Granovetter, who was interested in conditions under which individuals can *acquire* information, we evaluate bots in terms of how far they manage to *spread* their content. A second important difference is that Granovetter analyzed static networks. The strength of weak bots, in contrast, emerges because weak bots have a different impact on the diffusion dynamics in the network. Third, seemingly weak bots are strong, because their contacts do not grow too dissimilar to their friends and, thus, keep sufficiently strong ties to these friends and manage to spread bot content. Thus, weak bots are strong because they maintain strong rather than weak ties between the bot contacts and their friends.

3. Model

In order to demonstrate the effectiveness of seemingly ineffective bots, we build upon Axelrod's model for the dissemination of culture [19], one of the most influential models of social influence in networks [24]. In this model, agents are described by a set of features and exert influence on the feature set of their network neighbors. In his seminal paper, Axelrod studied the conditions under which repeated social influence leads to the emergence of consensus or polarized feature distributions with subgroups disagreeing on all features. This model has been widely adopted and has already been used to derive testable hypotheses about various phenomena, including the polarization of political opinions [25], mass media influence [26], or opinion dynamics in online social networks [27]. Here, we adopt Axelrod's model and add a bot who holds a fixed set of features and communicates them to users connected to the bot. We test whether the number of ties our simulated bots have to users and the activity of the bots in terms of the relative frequency of emitted messages affect the number of agents adopting features introduced by the bot. We keep all model assumptions that are not related to bot behavior unchanged, as Axelrod's model is very well understood, an aspect that makes it easier to demonstrate why the model generates the counter-intuitive effects of bots [28].

Our implementation of bots resembles earlier work by social-influence modelers studying the effects of charismatic leaders and extremists on opinion dynamics with the bounded-confidence model [16, 17]. For two reasons, however, we study effects in Axelrod's modeling framework rather than the bounded-confidence model. First, studying a different modeling framework allows us to explore the robustness of earlier findings to changes in model assumptions and to test whether earlier findings might hinge on characteristics of the modeling framework. Second, we deem Axelrod's framework more suitable for the study of bots in online social networks. In the bounded-confidence model, agents are described by opinions measured on a continuous scale. When an extremist or a bot exerts influence on an agent, this agent's opinion shifts closer to the opinion of the bot. Subsequently, the agent can influence other agents' opinions. As the agent's opinion shifted towards the bot, the bot exerts an indirect influence on these other agents, but this influence is moderated by the agent. In the context of online social networks, however, bots emit content that users can share with their contacts. As a consequence, the content sent by the bot is diffusing in the network without mediation. In contrast to the bounded-confidence model, this diffusion is directly represented in Axelrod's model where bots communicate beliefs that agents can adopt and forward to their contacts.

Adopting Axelrod's model, we generated N agents who each hold F beliefs about the world.[2] These beliefs are nominal characteristics with Q possible traits per belief. For instance, one of the F belief dimensions could represent different theories of the origin of the coronavirus. The traits could represent (1) that the coronavirus has a zoonotic origin, (2) that it has been genetically engineered in a CIA weapon program, or (3) that it has been stolen from a Canadian virus research laboratory. At the outset of a simulation run, all agent beliefs are initialized to a random value $q \in \{0, \dots, Q\}$ drawn with equal probability $(1/Q)$. All agent's beliefs are stored in the matrix C:

$$C = \left\{ \begin{matrix} q_{11} & q_{12} & \cdots & q_{1F} \\ q_{21} & q_{22} & \cdots & q_{2F} \\ \vdots & \vdots & \ddots & \vdots \\ q_{N1} & q_{N2} & \cdots & q_{NF} \end{matrix} \right\}$$

Agent-based models of social influence typically represent the beliefs of agents as (a set of) nominal variables, continuous variables, or a combination of the two [24]. In many occasions, nominal and continuous implementations do not substantively change the model dynamics [24]. Here, we opt for a vector with nominal variables, because it makes the influence of the bot easily traceable in equilibrium. If we observe a trait in the agent's feature vector that had not been considered by any agent but the bot at the outset, we know that the bot successfully influenced this agent's beliefs. To test whether the choice for representing beliefs as nominal traits substantively changed model dynamics one can compare our results, at least qualitatively, to the results of Hegselmann and Krause, who implemented a similar model with a continuous opinion dimension [16].

In our model, agents are represented as nodes in a network with undirected network links.[3] A link between two agents represents their opportunity to interact and communicate beliefs. That is, connected agents send and receive messages communicating their beliefs to each

[2] Axelrod referred to these beliefs as "features"

[3] In Section 4.4.4 we investigate whether networks with directed links produce dynamics similar to undirected networks.

other. In our simulations, we studied ring networks. That is, we arranged agents on a ring and created network ties between every node and the k closest neighboring nodes on the ring. The resulting network is characterized by a high degree of clustering, as there are many so-called "triplets", sets of three connected agents. This mimics a central characteristic of online social networks, where friends of friends tend to be friends.[4]

The model's dynamics are broken down into a sequence of discrete events t. At each event, an agent i is randomly picked for emitting a message to one of its network neighbors j. Also agent j is picked randomly from the set of network neighbors of agent i. Next, agent i sends a message to j, communicating a belief where the two agents disagree. With a probability p_s equal to the overall belief similarity between i and j, agent j adopts the belief communicated in the message. With this assumption, Axelrod implemented homophily, the notion that individuals tend to interact mainly with like-minded others. Homophily is a strong force in human behavior [29,30] and is reinforced by personalized recommender systems installed in online social networks [15,31,32]. These systems rank higher incoming messages emitted by like-minded users and, thus, increase chances that users are reading these messages. Formally, the probability that agent j adopts the communicated belief equals the normalized inverted Hamming distance:

$$p_s = 1 - \frac{\sharp\left\{f : q_{if} \neq q_{jf}, f = 1, \ldots, F\right\}}{F}$$

To model the presence of a bot in the network, we added one additional bot-agent to each simulated network, who held F randomly picked beliefs. These beliefs were fixed to implement that the bot cannot be influenced by its contacts. What is more, one of the bots' beliefs adopted a value outside of the $[1, Q]$ range, which represents that the bot agent is communicating a foreign belief.[5] The remaining $Q - 1$ beliefs adopted values that also other agents could have adopted. Otherwise, the bot would be maximally different from the remaining agents and, thus, unable to exert influence on others. The degree to which the foreign belief is adopted by the remaining agents in the network is the central outcome variable of our analyses. Conceptually, the bots in this model are exactly the same as the other agents except that their belief vector is immutable, and that the rate at which they are activated and emit content may vary.

We connected the bot-agent to a random subset of agents, unless specified otherwise (see Section 4.4.5). Parameter p_C allows influencing the proportion of agents who were connected to the bot and could, thus, receive messages from the bot. This parameter controls the "connectedness of the bot". Also, we added a parameter controlling the "activity of the bot", the probability p_A that in a simulation event the bot was emitting a message to one of its contacts. Experimental manipulation of the two parameters p_A and p_C allow us to test whether the model generates the counter-intuitive effect proposed above. That is, we tested whether a larger share of agents adopted the foreign belief when the bot agent had a low connectedness p_C and a low activity p_A.

Algorithm 1 details the steps of a simulation run. The model was implemented in python, using 'defSim', a software package specifically designed for discrete event social influence modeling [33]. A Jupyter notebook, and all the files used for the simulation experiments are available in the supplementary material.

Algorithm 1: Pseudo-code for a simulation run of the agent-based model.

```
initialize $F \times N$ matrix $C$ with random draws from $[1, ..., Q]$
initialize ring network where each agent is connected to $k$ nearest neighbors
create bot agent with $q = \{-1, 1, 1\}$
create links between bot agent and a random set of agents of size $p_C N$
set iteration = 0
while not all differences between neighbors are 0 or 1:
  if random float < $p_A$:
    agent $i$= bot
  else:
    agent $i$= one of $[1, ..., N]$
  agent $j$ = one of neighbors of agent $i$
  draw $f$ = one of $[1, ..., F]$
  if random float < $p_s$:
    agent $j$ copies trait $q_f$ of agent $i$
```

The model generates two main classes of equilibria, states where further sending and receiving of messages cannot change agents' beliefs. First, the population can develop a belief consensus in that all agents hold the exact same beliefs, a state that Axelrod referred to as "monoculture". Since the bot agent's beliefs are fixed, either all agents in the population adopted all F beliefs of the bot or none of the agents considers any of the bot beliefs. Second, it is possible that the network falls apart into mutually different but internally homogeneous segments. Within a segment, belief communication does not generate any change because agents already hold the same beliefs. Between the segments, there is no further communication of beliefs as connected agents belonging to two different segments consider different beliefs and, therefore, fail to further communicate beliefs. Axelrod called this a state of "polarization". The agents in one of the segments may have adopted all beliefs of the bot. All remaining segments containing agents with a network link to an agent in this segment, however, have to differ on all F belief dimensions. There may be segments in the network that have adopted only a subset of the bot's beliefs. However, no agent in such a segment has a direct network link to the bot or the segment that has adopted all bot beliefs.

4. Results

4.1. Equilibrium analysis

The central research question of this analysis is how social bots' network connectedness and messaging activity relate to their effectiveness in influencing the distribution of beliefs in the population. To answer this question, we conducted a simulation experiment, varying bot connectedness p_C and bot activity p_A from 0.05 to 0.95 in steps of 0.1. For each of the 100 experimental conditions, we conducted 25 independent simulation runs, assuming $N = 144$, $k = 12$, $F = 3$, $Q = 3$ in all runs. The main outcome variable is the *effectiveness of the bot*, measured as the share of agents having adopted the foreign bot-belief ($q_{bot,1}$) in equilibrium. In Section 4.2, we address the dynamics leading to equilibrium.

Fig. 1 informs about the share of agents having adopted the bot-belief in the whole population (left panel), among the bot's direct network neighbors (center panel), and among the agents who are not directly connected to the bot (right panel). Each cell in the heatmaps visualizes the average share observed in the 25 simulations per experimental condition. Darker shades of blue, red, and green visualize higher average bot effectiveness.

Panel A of Fig. 1 clearly refutes the naive intuition that more active and connected bots are more harmful. On average, fewer and not more agents adopted the bot-belief when the bot was more active and more embedded in the network.

[4] Originally, Axelrod studied a cellular automaton. Here, we opt for ring networks with a higher degree of clustering, as this allowed us to also study the effects of network structure on the strength of weak bots (see Section 4.4.2). The effects of this modeling choice are tested and discussed in the sensitivity analyses of Sections 4.4.2 and 4.4.3.

[5] This affects the ex-ante average similarity between the bot and his neighbors. Instead of adjusting for this difference, we chose to keep this assumption, as the estimated effect sizes of bot influence remain on the conservative side.

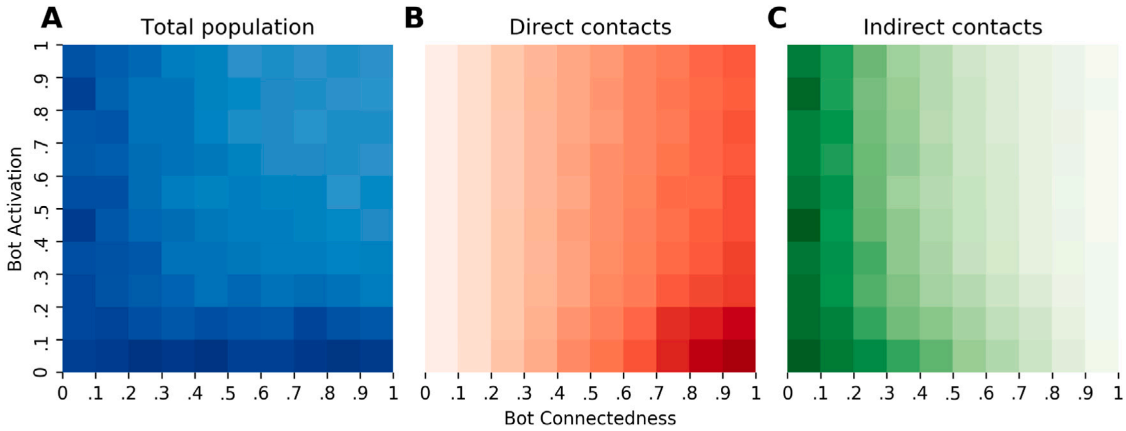

Fig. 1. Effectiveness of bots depending on connectedness and activity in equilibrium. Panel **A** shows the share of agents who adopted the bot belief in the whole population. Panel **B** depicts the share of agents directly connected the bot who adopted the bot belief and Panel **C** shows share of agents without a direct link to the bot who adopted the bot belief. Colors correspond to average shares in equilibrium over 25 independent simulation runs. Darker cells identify higher shares. Exact values are reported in the supplementary material.

The story of this seemingly counter-intuitive link between bot connectedness and effectiveness is a story of *in*direct influence. By persuading its direct contacts, bots exert indirect influence on agents connected to one or more of the bot's direct contacts. Panels B and C of Fig. 1 show that connectedness did make the bots more effective in convincing its direct neighbors (e.g., moving from 5% to 73% average share at $p_A = .15$), but at the cost of reducing the proportion of indirect contacts reached with the belief (from 79% to 4%). Note that bot connectedness increases the absolute number of neighbors adopting the belief (a simple consequence of opportunity), but decreases the share of persuaded direct neighbors. What is more, the increase in the absolute number of persuaded bot neighbors does not compensate for the much stronger negative effect of connectedness on bot effectiveness among indirect neighbors, composing a net negative effect. This finding is in line with Conjecture 1, formulated in Section 2: highly connected bots are not stronger than weakly connected bots.

Bot activity p_A appears to have similar effects as p_C (see Panel A of Fig. 1), but its link to direct and indirect influence is somewhat less obvious. Panels B and C of Fig. 1 do show a moderate negative effect at most levels of bot connectedness, but the transition is less clear than in the former case. Most notably, the effect of increasing activity on the bot's direct contacts is most pronounced at high levels of connectedness (e.g. from 84% to 55% at $p_C = .95$). Under this condition, a less active bot is more effective at persuading its direct contacts through indirect influence. It allows for relatively more interaction between the other agents, leading to a higher share of dissemination of its unique trait in equilibrium. As such, a spillover effect of indirect influence leads to the surprising finding that more active bots are less effective at persuading even their direct contacts. This supersedes conjecture 2, formulated in Section 2.

In simulation runs with a highly connected and very active bot, the population can quickly fall apart into segments consisting of agents who are either very similar or very dissimilar to the bot. As most interaction between non-bot agents happens inside of the segments, each segment grows increasingly homogeneous. As a consequence, communication between segments breaks down. When the bot, however, communicates its beliefs less actively, these segments do not form, and there is more communication between agents. In these communication events, bot beliefs can diffuse in the network and can reach agents who had grown too dissimilar to the bot already. Through indirect influence the bot now reaches even those agents who were too dissimilar at the outset to be influenced by the bot directly.

The variance of bot effectiveness within each experimental condition of the simulation experiment revealed another interesting pattern (figures are provided in the supplementary materials). We observed that the darker cells in all panels of Fig. 1 correspond to a more substantial variance in the outcomes of simulation runs. This is not trivial, since averages in those cells are closer to the maximum, and hence one would expect variances to decrease. Higher variance means that it is harder to predict the trajectory of a piece of (mis)information, which resonates with the empirical pattern that the content of weakly connected bots occasionally happens to successfully penetrate public debate.

4.2. Analysis of model dynamics

To illustrate the model's dynamics, we describe in this section ideal-typical simulation runs with weak and strong bots. In Fig. 2 we show trajectories of runs with a bot who is highly active and weakly connected (Panels A), weakly active and highly connected (Panels B), and one with a weakly active and weakly connected bot (Panels C).[6] Panels in the top row show the share of agents who have adopted the belief unique to the bot (Panels a), and panels in the bottom row show the absolute number of agents who adopted the trait (Panels b) and the cumulative distribution of belief change inflicted by the bot and the other agents.[7]

Panels A.a and A.b of Fig. 2 show typical dynamics generated by a highly active social bot with low connectedness. Since this bot was very actively communicating to a relatively small number of direct neighbors, the bot trait was spreading very successfully amongst the direct neighbors. Agents indirectly connected to the bot, however, adopted the belief at a much slower pace. Panel A.c, where the number of successful interactions is plotted over time, shows that after the first phase, in which the bot convinced its direct contacts, fewer interactions were successful (see red line). This indicates that agents unconnected to the bot refused to be influenced by bot neighbors because they had grown too dissimilar. The remaining successful interactions happened between agents who were not connected to the bot or direct neighbors of the bot who had adopted bot beliefs. Dynamics reached equilibrium when these agents developed a local consensus.

[6] The combination high activity and high connectedness is presented in the supplementary material for concision.

[7] Note that the time scale is hard to compare across the three runs, as high activity implies that a larger share of the simulation events was used by the bot rather than for communication between agents. Furthermore, an increased connectedness implies that it will take the bot a higher number of simulation events to reach all of its network neighbors because only one of them can be reached per event.

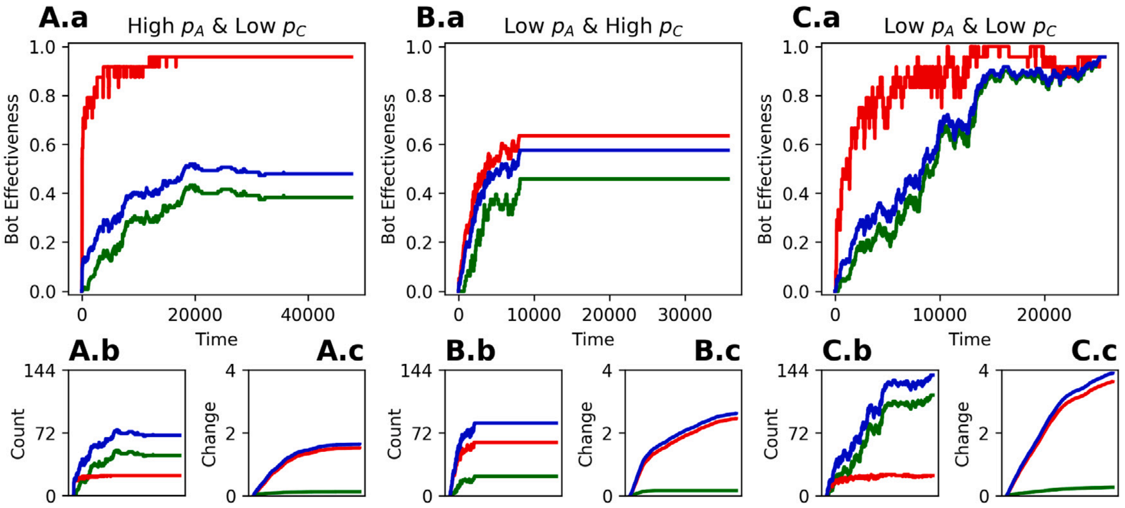

Fig. 2. Model dynamics in three typical runs; one with a highly active and weakly connected bot (panel **A**), one with a moderately active and highly connected bot (panel **B**), and one with a moderately active and weakly connected bot (panel **C**). 'High' p_C or $p_A = 1/6$, 'low' p_C or $p_A = 2/3$. Sub-panels **a** show trajectories of share agents who have adopted the bot trait in the whole population (blue), among the bot's direct network neighbors (red) and indirect network contacts (green). Sub-panels **b** show the same information, but in absolute numbers of agents. Sub-panels **c** show the cumulative distribution of successful influence events (blue) by non-bot agents (red) and by the bot (green). An interaction between two selected agents i and j is considered successful when j adopted a trait from i. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Panels B of Fig. 2, describe a typical run with low bot activity and high bot connectivity. Thus, this bot had many network neighbors but communicated infrequently. Compared to the run shown in Panels A, this bot did a bad job in convincing its direct network contacts, which is not surprising as the bot was not very active. As a consequence, it is also not surprising that this bot influenced relatively few indirectly connected agents. The problem was that many bot neighbors did not adopt the bot's beliefs and, at some moment, grew too dissimilar to interact with the bot. As a consequence, the bot did not communicate successfully any longer (see green line in Panel B.c). The direct and indirect neighbors of the bot who had not adopted the bot beliefs developed a consensus on beliefs that the bot did not share.

The bot shown in Panels C was the least connected and least active of the three, but it was most successful. This bot managed to steadily increase the share of direct neighbors who adopted the bot trait. However, the low bot activity made sure that the bot's neighbors did not adopt all bot beliefs and, thus, always kept beliefs shared with their other neighbors. As a consequence, the bot's direct neighbors managed to communicate bot beliefs to their contacts.

4.3. Statistical analysis of relationships

Fig. 1 showed that both bot activity and bot connectedness made bots less successful. The figure, however, does not reveal the precise strength of the two effects and whether they might strengthen or weaken each other. To explore in more detail the effects of bot activity and bot connectedness on the share of agents who adopted the bot belief in equilibrium, we conducted a regression analysis of the data from the main simulation experiment described in Section 4.1 [34].[8]

Table 1 shows the results from three regression models, with the share of the population that had adopted the bot belief in equilibrium as the dependent variable. All regression coefficients are statistically significant, showing that we conducted a sufficient number of replications per experimental treatment condition. Model 1 contains only the intercept and the two main effects of the experimental treatments. Model 2 adds squared terms, and Model 3, in addition, contains an interaction effect. All regression models indeed display the negative relationship between bot connectedness, bot activity, and bot effectiveness, supporting the strength-of-weak-bots effect again.

Table 1
Ordinary least squares regression model on the diffusion of the bot trait in equilibrium.

	Model 1	Model 2	Model 3
Parameters			
Intercept	0.918	1.050	0.991
Connectedness	−.191	−0.476	−0.357
Activity	−0.234	−0.737	−0.618
Connectedness^2		0.285	0.285
Activity^2		0.503	0.503
Connectedness × Activity			−0.238
Fit statistics			
R^2	0.335	0.413	0.430
AIC	−3407	−3717	−3789

Furthermore, Model 2 suggests that this negative effect of both connectedness and activation is reduced as they move closer to 1, implying non-linear relationships. Their interaction, however, strengthens both effects as Model 3 shows. The AIC's of the models suggest that the most expansive one fits best, although the improvement from Model 3 upon Model 2 is marginal as the small increase in explained variance shows. Moreover, the results from all models should be taken with a grain of salt, as we have seen that the variance between all levels of the independent variables differs substantively, violating the homoscedasticity assumption.

4.4. Sensitivity analyses

While Axerod's model is a fruitful point of departure for the study of bot effects, Axelrod made a series of assumptions that do not resonate with the context of online social networks. In order to test the robustness of our results to changes in potentially important model assumptions, we conducted a series of sensitivity analyses. However, not every model assumption that seems to deviate from reality has the

[8] Agent based models do not require statistical analysis since the indicators are not estimations underlying data generating processes, but reflect actual realizations of an artificial process. Nevertheless, regression analysis allows one to describe parameter's effects and their interdependencies with great precision [34]. In particular, a regression model can uncover the relative effects of the parameters of interest, provide insights into model sensitivity, and present complex relationships in a familiar, easy to interpret manner.

potential to affect model predictions about bot effectiveness. Accordingly, we focused our sensitivity tests on assumptions where earlier work with Axelrod's model found effects on the dynamics of consensus formation and polarization and tested whether or not these assumptions also affect the strength of weak bots. In particular, we explored model assumptions about the communication regime (Section 4.4.1), network structure (Sections 4.4.2–4.4.4), and node heterogeneity (Section 4.4.5). All codes used and more detailed results are available in the supplementary material.

4.4.1. One-to-one vs. one-to-many communication

So far, we adopted Axelrod's original model of the dissemination of culture as closely as possible. While this makes our findings directly comparable to earlier work, a possible downside of our approach is that some of Axelrod's assumptions may be problematic when applied to the context of online social networks. Our earlier work [27], for instance, has shown that model dynamics can change drastically when the model is tailored to the communication regime of online social networks. While Axelrod used one-to-one communication between agents, communication in online social networks is better described by a one-to-many regime, because users of these systems tend to share messages with all of their connections at the same time. We showed that communicating to many neighbors at once generates more cultural polarization [27]. We tested here whether the strength-of-weak-bots effect is affected by changing the communication regime from a one-to-one to a one-to-many world. To this end, we conducted a second simulation experiment, where we compared the two communication regimes. We kept bot connectedness constant at $p_C = 1/3$ (48 of 144 agents), and varied bot activity: $p_A = x/144$ for $x \in \{1, 2, 3, 6, 12, 24, 48, 96\}$. All codes used and more detailed results are available in the supplementary material.

Fig. 3 shows that we found very similar results for both communication regimes, suggesting that our findings are robust to changes in the communication regime. If anything, the effect is even more pronounced under one-to-many communication.

4.4.2. The effect of network clustering

So far, we conducted all analyses with networks that are characterized by high network clustering. In these networks, agents tend to be connected to agents who are also directly linked, representing the notion that "friends of friends tend to be friends". While empirical research showed that online social networks tend to be highly clustered [35], it is also known that network clustering has a strong effect on the diffusion of traits in the network [e.g. 36,37]. When, as assumed in Axelrod's model, agents adopt a trait after having been exposed to it by a single source (so-called "simple contagion"), network clustering hampers diffusion, as in clustered networks many ties are redundant for the diffusion. That is, these ties create connections between nodes that are not contributing to the diffusion because other ties have established a connection already. This suggests that network clustering makes it more difficult for bots to spread beliefs, ceteris paribus. However, the redundancies present in highly clustered networks might also decrease the weakness of strong bots, since each of them provides bots with an additional path to the neighbors of their neighbors. As a consequence, the strength-of-weak-bots effect may be weaker in clustered networks.

In order to test this conjecture, we conducted additional simulations, experimentally manipulating the number of network ties that were randomly rewired. That is, we generated the described ring networks and randomly rewired a share of the ties. We rewired a share $p_R \in \{0, .01, .02, .04, .08, .16, .32, .64, 1\}$. The lowest rewiring probability ($p_R = 0$) generates the same ring networks as studied above. When the highest value ($p_R = 1$) is implemented, all network ties of the ring network are replaced by a link between two randomly picked agents. Bot connectedness was set to $p_C = 1/3$ and bot activity was either low ($p_A = 1/24$) or high ($p_A = 1/3$).

Fig. 4 reveals two main findings. First, when more links had been rewired (that is, clustering is decreased), more agents adopted the bot-belief. This replicates the mentioned effect that network clustering hampers the diffusion of traits in networks. Second, the blue lines are consistently below the red lines, which shows that increased bot activity makes bots less effective in convincing direct and indirect network contacts. This effect is found for all studied network structures, independent of their degree of clustering. Thus, the strength-of-weak-bots effect is robust to changes in network clustering.

4.4.3. Lattice networks

So far, we focused our analysis on a very simple network structure, a ring network. While this network structure shares essential characteristics with real networks (in particular, high clustering), it also has characteristics that may affect the diffusion of beliefs in a network. On a ring network, in particular, a belief can diffuse only in two directions on the network, clockwise and counter-clockwise. To test whether our findings may depend on this aspect, we tested whether results change when a lattice network is implemented. On a lattice network, agents are not arranged on a circle and connected to the closest k neighbors. Instead, agents are arranged on a grid. As a consequence, agents do not only have ties to the right and the left, but they have connections in all directions. As a consequence, a belief can also spread into a higher number of directions, which could amplify the diffusion of false beliefs. It is unclear whether this affects the effectiveness of bots.

To test whether our findings hinge on the assumption of ring networks, we implemented the lattice-network structure that also Axelrod assumed in his seminal paper [19]. There were no boundary conditions. In addition, we varied the size of agents' neighborhood between the typical "Moore" neighborhood (where each agent is connected to 8 of its nearest neighbors in a square), and the "Von Neumann" neighborhood (where each agent links to 4 agents on the adjacent squares). We set $p_C = 1/3$, and $p_A = x/144$ for $x \in \{1, 2, 3, 6, 12, 24, 48, 96\}$. The bots were implemented in the same way as described above, being linked to a random share of one third of the population.

Fig. 5 shows for both lattice networks that bot activity decreased the bot's effect on the whole population, its direct network neighbors, and all agents indirectly connected to the bot. Thus, we found the same patterns as in the ring networks, showing that these findings are robust. The effect's difference between the two neighborhood conditions turned out to be very small. Only in the condition with a minimal bot activity, there is a consistent and significant difference between the two conditions. Future research is needed to explain this observation.

4.4.4. Directed networks

While many online social media platforms were originally designed as peer-to-peer platforms with non-directed links between users, there are nowadays plenty of platforms where users establish directed relationships. Most notably, on Twitter and Instagram users form directed connections when they "follow" another account. Such directed connections are often reciprocated, but in particular nodes with high numbers of followers fail to reciprocate. The notion that link directionality could impact the dynamics of beliefs in influence networks is not new [38], but it is hard to anticipate whether the strength-of-weak-bots effect is affected by network directionality. One the one hand, the social mechanism that makes weak bots strong (weak bots' friends remain sufficiently similar their friends to spread the bot's message) should not depend on whether links are directed or not. On the other hand, directed links might weaken the effect, all other things being equal. When a user received a bot trait from an account she follows, she cannot communicate the trait back to the account, which increases the chance that the person drops the trait at a later point in time.

Comparing dynamics in directed and undirected networks is challenging, as it is not possible to generate a directed and an undirected network without changing other potentially critical characteristics of

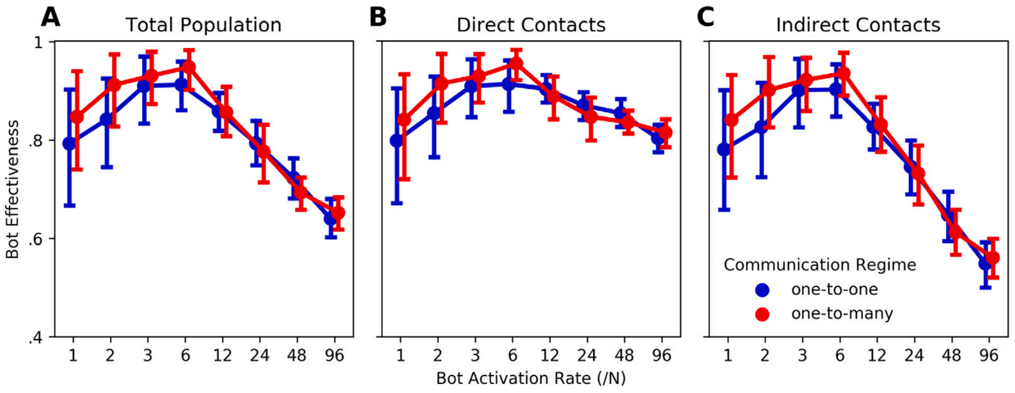

Fig. 3. Comparison of communication regimes in terms of bot effectiveness. Lines show share of agents who adopted the bot belief in the whole population (Panel **A**), amongst agents with a direct link to the bot (Panel **B**), and amongst agents without a direct link to the bot (Panel **C**), averaged over 25 independent simulation runs per condition. Error bars depict the 95% confidence interval. Note that the lower end of the y-axis is cut at .40 for visual clarity.

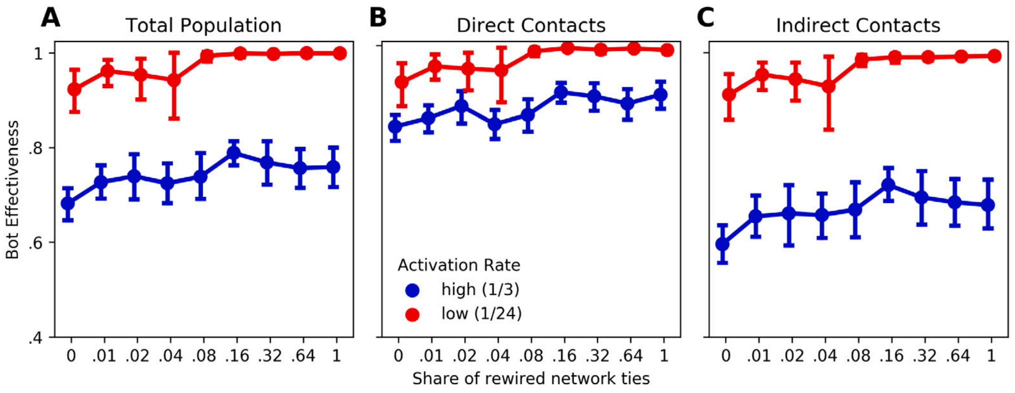

Fig. 4. Effect of network clustering on bot effectiveness measured as the share of agents who adopted the bot belief in the whole population (Panel **A**), amongst agents with a direct link to the bot (Panel **B**), and amongst agents without a direct link to the bot (Panel **C**), averaged over 25 independent simulation runs per condition. Error bars depict the 95% confidence interval. Higher shares of rewired network ties translate into weaker network clustering.

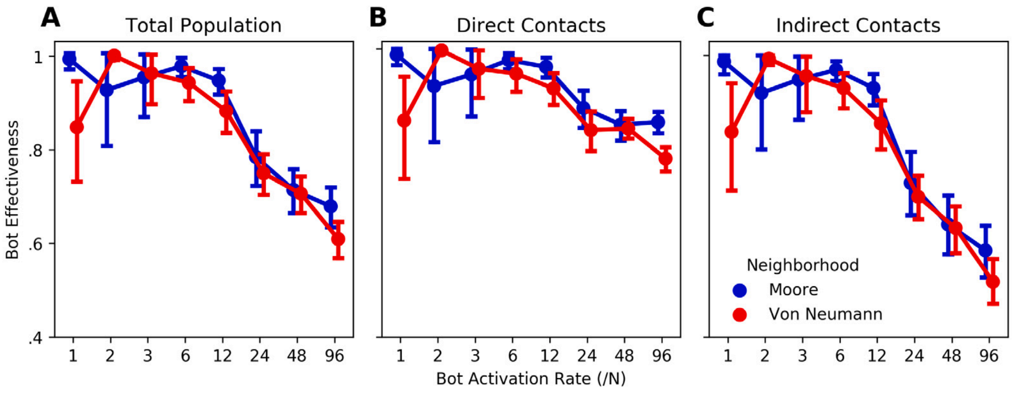

Fig. 5. Replication of the strength-of-weak-bots effect on lattice networks. Lines show share of agents who adopted the bot belief in the whole population (Panel **A**), amongst agents with a direct link to the bot (Panel **B**), and amongst agents without a direct link to the bot (Panel **C**), averaged over 25 independent simulation runs per condition. Error bars depict the 95% confidence interval.

the network. One could, for instance, start from a directed network and turn one undirected link into two directed ties. The resulting network would be identical and dynamics would not change. Erasing one of the directed link would imply a change in the network, but it would also alter the degree distribution in the network. Likewise, keeping the link and rewiring it would increase the number of incoming ties for one other node, also changing the degree distribution. As a consequence, it is difficult to attribute differences between bot effects observed in directed and undirected networks to directionality, as also other characteristics of the network have changed.

Thus, rather than comparing bot effects in directed and undirected networks, we tested whether it is possible to replicate our main findings in an undirected network. To this end, we generated a series of spatial random graphs [39] with 144 nodes, each with an outdegree of 12 ($k = 12$). This network generator does allow modeling directed ties, whilst keeping network clustering and the degree distribution comparable to the networks studied above. Spatial random graphs are considered realistic representations of human social networks because they share important characteristics such as a high level of clustering, low tie density, short average geodesic distance, and a community structure.

The networks we created have an observed average reciprocity value of 81% and transitivity value of 61%. We conducted a simulation experiment, studying four treatments with high and low bot connectedness, as well as high and low bot activity (p_C and $p_A \in \{\frac{1}{12}, \frac{1}{2}\}$) and conducted 25 independent replications per treatment. Fig. 6 reveals that the strength-of-weak-bots effects is also present in directed networks.

4.4.5. Unbalanced degree distributions

There is anecdotal evidence suggesting that so-called "*influencers*", human users with a high number of followers, can amplify the spreading of fake news-stories through successive posting [40]. So far, however, our analyses focused on networks where all human users have the same degree (number of network ties). Therefore, we tested whether our findings also obtain in networks where degree varies and whether human influencers might interact with the strength-of-weak-bots effect.

To this end, we implemented the spatial random graphs used in the previous Section 4.4.4, but this time we assigned each agent an outdegree k_i drawn from a Poisson distribution with an average of 12. What is more, we devised a quasi-experimental test for three competing scenarios. That is, we connected the bot to the proportion p_C of (i) those other agents who had the highest outdegree in the network (the influencers), (ii) those other agents who had the lowest outdegree in the network, or (iii) randomly picked agents. In this way, we tested whether influencers have the potential to aid bots that fail to reach larger shares of human users. For each of the three competing scenario's, we ran a simulation experiment with four conditions of high and low bot connectedness and low bot activity (p_C and $p_A \in \{\frac{1}{12}, \frac{1}{2}\}$). All twelve distinct conditions were replicated 25 times.

Fig. 7 confirms the pattern that we observed earlier. When the bot is connected to a random subset of the population (Panels A of Fig. 7), its activity and connectivity rates are both negatively related to the proportion of the population that adopted the bot's unique trait. Using the other bot connection procedures does not seem to disturb the strength-of-weak-bots effect at high rates of connectivity. There (in the higher panels of Fig. 7), the effect of bot activity is strikingly similar to the random matching procedure. Remarkably, the bot is as effective at persuading the population at large when it is linked to the least well connected half of the population, as when it is linked to the best connected half of all agents. At lower levels of bot connectivity, however, there is an interesting qualitative difference of the effect of bot activity. When the bot is connected to the agents with the highest outdegree, it seems slightly more apt to persuade all others in the network than the bots in the random matching situation. Nevertheless, the (small) negative effect of bot activity appears to remain. When the bot is connected to the one twelfth share of agents with the lowest outdegree, the relationship between bot activity and bot effectiveness is flipped. A bot trying to influence a group of agents in the periphery of the network, with only a small number of contacts, can still effectively influence the population at large. What is more, the difference may be small, but the bot is more effective at persuading the larger population than when it would be connected to the one twelfth share of agents with the highest outdegree. An explanation for this surprising finding could be that agents with a low outdegree, are likely to be embedded in peripheral parts of the network with a high degree of local clustering. The spatial random graph favors the creation of ties with agents that are close, to achieve clustered graphs with a community structure. Those smaller subsets may be excellent breeding grounds for ideological similarity because they offer more opportunity for reinforcement than fragmented graphs.

5. Summary and discussion

Social bots have been identified as a potential threat to public opinion formation and democratic decision-making. Empirical research on bots has led to seemingly inconsistent results, showing that, on the one hand, bots tend to have contact to a small number of human users and that, on the other hand, the content that bots spread can reach and influence large parts of online social networks. In this paper, we proposed a theoretical explanation reconciling these seemingly contradictory findings, arguing that bots do not effectively spread (false) beliefs *despite* but *because* of their limited effectiveness in convincing directly connected users. Bots with direct influence on a small number of users and limited activity can exert a stronger influence on the whole population, because their direct contacts are influenced slowly and, therefore, keep communicating with their network neighbors. As a consequence, a seemingly weak bot can exert indirect influence on a larger share of the population, and its messages reach more users.

Using an agent-based model of social influence, we showed that weakly connected and moderately active bots are more effective in spreading beliefs in the network at large. In a series of simulation experiments, we observed that a higher share of agents adopted beliefs communicated by a bot when the bot was exerting direct influence on fewer agents and when the bot was emitting the belief infrequently. As expected, we found that agents who were not directly connected to the bot adopted the bot's belief with a smaller likelihood when the bot was more active. Unexpectedly, however, we found that the bots' direct network neighbors also adopted the belief with a lower probability when the bot was more active. We argue that this unexpected finding also results from the complexity of the social-influence dynamic. When the bot's belief is not adopted by the neighbors of a bot's direct contact then these neighbors will not remind the direct neighbor of the belief when they drop it. We tested the sensitivity of the strength-of-weak-bots effect to changes in central model assumptions. It turned out that the effect is robust when one implements one-to-many rather than one-to-one communication, when networks with different degrees of clustering are attacked by the bot, when lattice rather than ring networks are studied, and when networks contain unreciprocated ties. What is more, we modeled networks with variable degree distributions and tested whether bots would be more successful if they linked to the most influential nodes. Not only does the strength-of-weak-bots effect prevail in those situations, the bot turned out to be even less effective than bots linked to niche communities.

This study offers two main insights for policymakers and engineers of social media platforms. First, the strength-of-weak-bots effect suggests that sparsely connected bots are not innocent and may even have a stronger effect than well-connected bots. Likewise, the seemingly powerful effects of well-connected and active bots and other social network users may be more limited than expected. To be able to evaluate the potential impact of bots on public opinion formation and democratic decision-making, it is important to quantify bots' ability to reach also indirectly connected users. Second, our findings echo a warning of potential overconfidence in bot-detection efforts to date. While weak bots may be particularly effective, their detection is likely to be more difficult as they emit fewer signals revealing that they are automated [41,42].

The strength-of-weak-bots effect has implications for future research. On the one hand, the effect turns out to be robust, having been replicated under various conditions, and with both Axelrod's dissemination-of-culture model and the bounded-confidence model [16]. This strong robustness suggests that the strength-of-weak-bots effect is not an artifact generated by a specific model under certain conditions, but may also be active in real online social media. On the other hand, it is still a largely untested hypothesis. This study served as an illustration of a theoretical mechanism in complex networks, and aimed to carefully assess its internal validity by means of robustness testing. Yet, while it can explain findings of earlier empirical research that appeared inconsistent from the perspective of other theories, direct tests of the effect are missing. The extent to which the strength-of-weak-bots effect is observable and of significant impact on the dissemination of (mis)information in online social media platforms remains an open question.

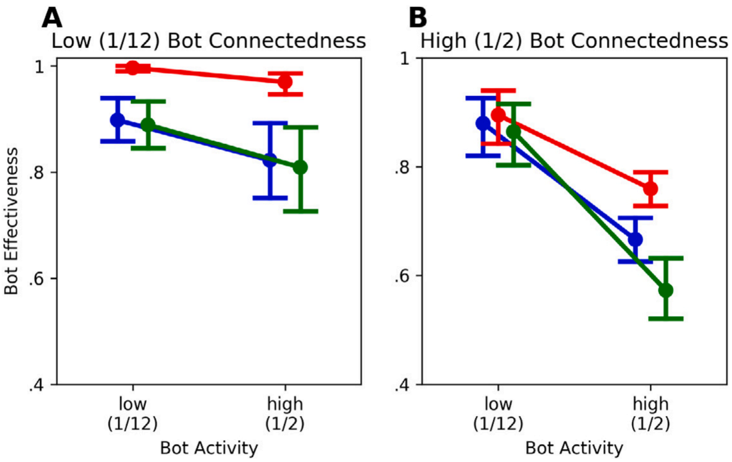

Fig. 6. Replication of the strength-of-weak-bots effects on directed networks. Dots show share of agents who adopted the bot belief in the whole population (**blue** dots), amongst agents with a direct link to the bot (**red** dots), and amongst agents without a direct link to the bot (**green** dots), averaged over 25 independent simulation runs per condition. Error bars depict the 95% confidence interval. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

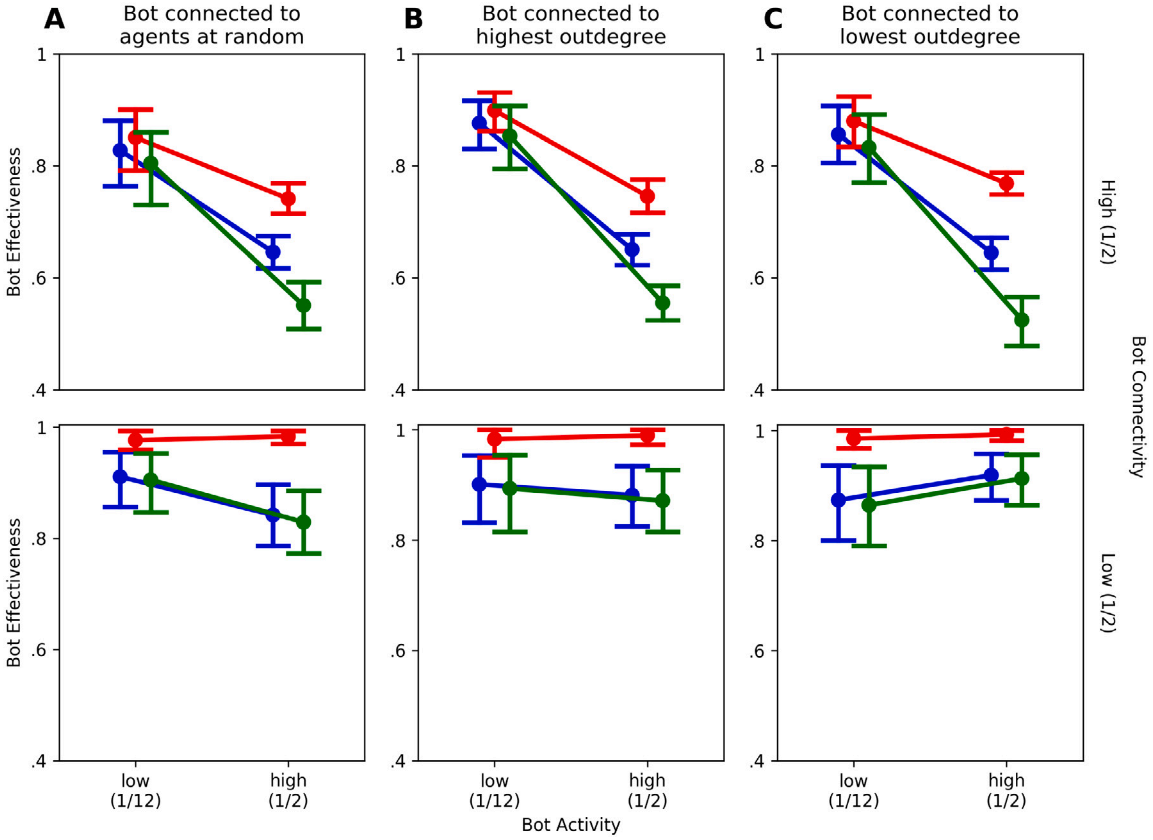

Fig. 7. Replication of the strength-of-weak-bots effects on networks where nodes have different numbers of network links. Dots show share of agents who adopted the bot belief in the whole population (**blue** dots), amongst agents with a direct link to the bot (**red** dots), and amongst agents without a direct link to the bot (**green** dots), averaged over 25 independent simulation runs per condition. Error bars depict the 95% confidence interval. Panel **A** shows scenario where the bot is connected to a random set of agents. In Panel **B**, the bot is connected to influencers, agents with many followers. In Panel **C**, the bot is connected to the agents with the smallest number of followers. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The strength-of-weak-bots effect, in addition, suggests that existing empirical research may have focused too much on bots' direct contexts [9]. While we did observe the effect also amongst bots' direct contacts, the main strength of weak bots results from their ability to reach users who are not directly connected to them, an aspect that deserves more empirical research.

In addition, empirical research is needed to test the unexpected model prediction that even the bot's direct network neighbors are eventually less affected by the bot when the bot is highly active. As we argued that a possible explanation for this finding is that the bot's direct neighbors may happen to drop the bot-emitted belief and may then not be reminded by their neighbors, as the belief never reached them, we recommend studying the sharing of bot content amongst human users empirically.

While our paper is motivated by the debate about the effectiveness of bots emitting false information, Hegselmann and Krause pointed to equally intriguing model implications for the dissemination of true beliefs [16]. Assume that the bot in our model is not a malicious program spreading falsehoods, but reality sending signals to agents seeking to identify the truth. These could be scientists who conduct studies and receive signals about the truth [43]. "However, if the truth seekers are 'too good' and converge too fast in the direction of the truth, they may leave behind them – and often far distant from the truth – major fractions of their not truth-seeking fellow citizens". [43, :505]. This suggests that future empirical research should not focus only on false information emitted by bots. According to the model, the counter-intuitive effects of communication activity and connectedness should be present also in the context of other forms of information and human users.

Seemingly weak bots can be strong because of the intermediate role of the bot's direct contacts. The model generates this effect without specific assumptions about characteristics of bots, content, users, or the communication context [11–14]. From a methodological perspective, this is very insightful, as it shows that the strength-weak-bots effect is an independent explanation. However, we do not argue that these other aspects are not relevant in real online social networks. As a consequence, when applied to real social networks, the empirical question emerges how strong the strength-of-weak-bots effect is relative to other factors. Empirical research answering this question is urgently needed. Furthermore, while we have spent most effort on understanding the mechanism and testing its robustness, there is still a lot unknown about the ways in which the strength-of-weak-bots effect can amplify or weaken other concurrent mechanisms. For example, a recent agent-based model that focused on social bot effects on the behavioral inclinations of agents to voice their opinions found that only a small number of bots is needed to create a spiral of silence that leads to over-representation of a (niche) opinion in a discussion network [44]. Taking the opinion positions and willingness to express these opinions into account at once, may amplify both the strength-of-weak-bots and spiral of silence effects.

Since the aim of the present analysis was to demonstrate a theoretical mechanism, we abstracted from many potentially important aspects. Bot behavior, content characteristics, network structure, or interpersonal influence can, of course, be formally captured in alternative ways. While we tested the robustness of some of our assumptions already in this paper, more work in this direction is needed. Earlier extensions of Axelrod's model demonstrated, for instance, that noise [e.g. 45], social (or 'many-to-one') influence [46], or 'globalization' through increased interaction range in larger networks [47] have the potential to change model predictions. More modeling work testing whether these aspects increase or decrease the strength-of-weak-bots effect is needed. This work will help develop testable hypotheses about the conditions under which bots are weak or strong and which parts of a social network graph are most resistant to bot attacks.

Supplementary material

All codes and data used are available in an online GitHub repository, at https://github.com/marijnkeijzer/weakbots/.

CRediT authorship contribution statement

Marijn A. Keijzer: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing, Visualization. **Michael Mäs:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank the Norms and Networks Cluster at the University of Groningen, in particular Andreas Flache, Dieko Bakker, Carlos de Matos Fernandes and Vincenz Frey, for their helpful feedback on an early version of this work. We are also grateful to the two anonymous reviewers who provided us with valuable feedback and suggestions. We would like to thank the Center for Information Technology of the University of Groningen, The Netherlands for their support and for providing access to the Peregrine high performance computing cluster.

References

[1] D. Ruths, Social science: The misinformation machine, Science 363 (6425) (2019) 348.

[2] A. Bessi, E. Ferrara, Social bots distort the 2016 U.S. presidental election, First Monday 21 (11) (2016) 1–15.

[3] S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online, Science 359 (March) (2018) 1146–1151.

[4] C. Shao, G.L. Ciampaglia, O. Varol, K.C. Yang, A. Flammini, F. Menczer, The spread of low-credibility content by social bots, Nat. Commun. 9 (1) (2018) 4787.

[5] R. DiResta, K. Shaffer, B. Ruppel, D. Sullivan, R. Matney, R. Fox, J. Albright, B. Johnson, The Tactics and Tropes of The Internet Research Agency, U.S. Senate Documents - Congress of the United States, 2018, p. 101.

[6] D.M.J. Lazer, M.A. Baum, Y. Benkler, A.J. Berinsky, K.M. Greenhill, F. Menczer, M.J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, M. Schudson, S.A. Sloman, C.R. Sunstein, E.A. Thorson, D.J. Watts, J.L. Zittrain, The science of fake news, Science 359 (6380) (2018) 1094–1096.

[7] M. Gentzkow, Small media, big impact, Science 358 (6364) (2017) 726–727.

[8] N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, D.M.J. Lazer, Fake news on Twitter during the 2016 U.S. presidential election, Science 363 (January) (2019) 6.

[9] C.A. Bail, B. Guay, E. Maloney, A. Combs, D.S. Hillygus, Assessing the Russian internet research agency ' s impact on the political attitudes and behaviors of American Twitter users in late 2017, Proc. Natl. Acad. Sci. (2019) 1–8.

[10] S. González-Bailón, M. De Domenico, Bots are less central than verified accounts during contentious political events, SSRN Electron. J. (2020) Available at: http://dx.doi.org/10.2139/ssrn.3637121.

[11] G. Pennycook, D.G. Rand, Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking, J. Pers. 88 (2) (2020) 185–200.

[12] W.J. Brady, J.A. Wills, J.T. Jost, J.A. Tucker, J.J. Van Bavel, Emotion shapes the diffusion of moralized content in social networks, Proc. Natl. Acad. Sci. 114 (28) (2017) 7313–7318.

[13] A. Guess, J. Nagler, J. Tucker, Less than you think: Prevalence and predictors of fake news dissemination on Facebook, Sci. Adv. (2019) 1–8.

[14] G. Pennycook, D.G. Rand, Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning, Cognition 188 (September 2017) (2019) 39–50.

[15] M.A. Keijzer, M. Mäs, The complex link between filter bubbles and opinion polarization, 2020, Available at: https://datasciencehub.net/system/files/ds-paper-629.pdf.

[16] R. Hegselmann, U. Krause, Opinion dynamics under the influence of radical groups, charismatic leaders, and other constant signals: A simple unifying model, Netw. Heterog. Media 10 (3) (2015) 477–509.

[17] J.D. Mathias, S. Huet, G. Deffuant, Bounded confidence model with fixed uncertainties and extremists: The opinions can keep fluctuating indefinitely, J. Artif. Soc. Soc. Simul. 19 (1) (2016) 1–15.

[18] M.W. Macy, A. Flache, Social dynamics from the bottom up: Agent-based models of social interaction, in: P. Hedström, P. Bearman (Eds.), The Oxford Handbook of Analytical Sociology, Oxford University Press, 2009, pp. 245–268.

[19] R.M. Axelrod, The dissemination of culture: A model with local convergence and global polarization, J. Confl. Resolut. 41 (2) (1997) 203–226.

[20] O. Varol, E. Ferrara, C.A. Davis, F. Menczer, A. Flammini, Online human-bot interactions: Detection, estimation, and characterization, in: Proceedings of the Eleventh International AAAI Conference on Web and Social Media, 2017, pp. 280–289.

[21] M. Mäs, The complexity perspective on the sociological micro-macro-problem, 2018, Available at SSRN: https://ssrn.com/abstract=3129362.

[22] E. Bakshy, J.M. Hofman, W.A. Mason, D.J. Watts, Everyone's an influencer: quantifying influence on twitter, in: Proceedings of the fourth ACM international conference on Web search and data mining SE - WSDM '11, 2011, pp. 65–74.

[23] M.S. Granovetter, The strength of weak ties, Am. J. Sociol. 78 (6) (1973) 1360–1380.

[24] A. Flache, M. Mäs, T. Feliciani, E. Chattoe-Brown, G. Deffuant, S. Huet, J. Lorenz, Models of social influence: towards the next frontiers, J. Artif. Soc. Soc. Simul. 20 (4) (2017).

[25] M.W. Macy, J.A. Kitts, A. Flache, S. Benard, Polarization in dynamic networks: A hopfield model of emergent structure, in: R. Breiger, K. Carley, P. Pattison (Eds.), Dynamic Social Network Modeling and Analysis, The National Academies Press, Washington, 2003, pp. 162–173.

[26] R. Ulloa, C. Kacperski, F. Sancho, Institutions and cultural diversity: Effects of democratic and propaganda processes on local convergence and global diversity, PLoS One 11 (4) (2016).

[27] M.A. Keijzer, M. Mäs, A. Flache, Communication in online social networks fosters cultural isolation, Complexity (2018) 1–20.

[28] R. Axtell, R.M. Axelrod, J.M. Epstein, M.D. Cohen, Aligning simulation models: a case study and results, Comput. Math. Organ. Theory 1 (2) (1996) 123–141.

[29] M. McPherson, L. Smith-Lovin, J.M. Cook, Birds of a feather: Homophily in social networks, Annu. Rev. Sociol. 27 (1) (2001) 415–444.

[30] P.F. Lazarsfeld, R. Merton, Frienship as a social process: A substantive and methodological analysis, in: M. Berger, T. Abel, C.H. Page (Eds.), Freedom and Control in Modern Society, Van Nostrand, New York, 1954, pp. 18–66.

[31] E. Pariser, The Filter Bubble: What the Internet Is Hiding from You, Penguin, London, United Kingdom, 2011, p. 304.

[32] A. Bruns, Are Filter Bubbles Real? John Wiley & Sons, 2019, p. 144.

[33] A.L. Laukemper, M.A. Keijzer, D.M. Bakker, defSim (v 0.1), GitHub, 2019, Retrieved from: https://github.com/marijnkeijzer/defSim/.

[34] A. Grow, Regression metamodels for sensitivity analysis in agent-based computational demography, in: A. Grow, J. Van Bavel (Eds.), Agent-Based Modelling in Population Studies, Springer, 2017, pp. 185–210.

[35] J. Ugander, B. Karrer, L. Backstrom, C. Marlow, The anatomy of the facebook social graph, 2011, pp. 1–17, Available at: http://arxiv.org/abs/1111.4503.

[36] R. Albert, H. Jeong, A.-L. Barabási, Error and attack tolerance of complex networks, Nature 406 (2000) 378–385.

[37] D. Centola, M.W. Macy, Complex contagions and the weakness of long ties, Am. J. Sociol. 113 (3) (2007) 702–734.

[38] J. French, Formal theory of social power, Psychol. Rev. (1956).

[39] L.H. Wong, P. Pattison, G. Robins, A spatial model for social networks, Physica A 360 (1) (2006) 99–120.

[40] R. Van Gool, C. Van de Ven, ‘Wij zijn het nieuwe nieuws’ [‘We are the new news’], De Groene Amsterdammer (2020).

[41] D. Assenmacher, L. Clever, L. Frischlich, T. Quandt, H. Trautmann, C. Grimme, Demystifying social bots: On the intelligence of automated social media actors, Soc. Media + Soc. 6 (3) (2020) 1–14.

[42] B. Kollanyi, Where do bots come from? An analysis of bot codes shared on github, Int. J. Commun. 10 (2016) 4932–4951.

[43] S. Balietti, M. Mäs, D. Helbing, On disciplinary fragmentation and scientific progress, PLoS One 10 (3) (2015) 1–26.

[44] B. Ross, L. Pilz, B. Cabrera, F. Brachten, G. Neubaum, S. Stieglitz, Are social bots a real threat? An agent-based model of the spiral of silence to analyse the impact of manipulative actors in social networks, Eur. J. Inf. Syst. 28 (4) (2019) 394–412.

[45] K. Klemm, V.M. Eguíluz, R. Toral, M. San Miguel, Nonequilibrium transitions in complex networks: A model of social interaction, Phys. Rev. E 67 (2) (2003) 026120.

[46] A. Flache, M.W. Macy, Local convergence and global diversity: From interpersonal to social influence, J. Confl. Resolut. 55 (6) (2011) 970–995.

[47] J.M. Greig, The end of geography? J. Confl. Resolut. 46 (2) (2002) 225–243.