Journal of Business Economics (2021) 91:1299–1329
https://doi.org/10.1007/s11573-020-01022-9

ORIGINAL PAPER

Diffusion and persistence of false rumors in social media networks: implications of searchability on rumor self-correction on Twitter

Kathrin Eismann[1]

Accepted: 2 December 2020 / Published online: 3 February 2021

Abstract

Social media networks (SMN) such as Facebook and Twitter are infamous for facilitating the spread of potentially false rumors. Although it has been argued that SMN enable their users to identify and challenge false rumors through collective efforts to make sense of unverified information—a process typically referred to as self-correction—evidence suggests that users frequently fail to distinguish among rumors before they have been resolved. How users evaluate the veracity of a rumor can depend on the appraisals of others who participate in a conversation. Affordances such as the searchability of SMN, which enables users to learn about a rumor through dedicated search and query features rather than relying on interactions with their relational connections, might therefore affect the veracity judgments at which they arrive. This paper uses agent-based simulations to illustrate that searchability can hinder actors seeking to evaluate the trustworthiness of a rumor's source and hence impede self-correction. The findings indicate that exchanges between related users can increase the likelihood that trustworthy agents transmit rumor messages, which can promote the propagation of useful information and corrective posts.

Keywords Affordances · Agent-based simulation and modelling · Sense-making · Social influence · Social media · Social networks

JEL Classification C63 · D79 · D83 · D85

✉ Kathrin Eismann
kathrin.eismann@uni-bamberg.de

[1] Department of Information Systems and Social Networks, University of Bamberg, Bamberg, Germany

1 Introduction

How rumors—that is, "item[s] of circulating information whose veracity status is yet to be verified at the time of posting" (Zubiaga et al. 2018, p 2)—diffuse and persist among the users of social media networks[1] (SMN) such as Facebook and Twitter is a timely subject of research. While rumors are, by definition, neither necessarily false nor harmful, they are often referred to in the same breath as fake news, hoaxes, and other forms of misinformation, and their impact on people's decisions is seen as potentially critical (Zannettou et al. 2019).

Despite their bad reputation, rumors may actually help people manage and make sense of situations they perceive as individually or collectively threatening (DiFonzo and Bordia 2007a). Still, rumors propagating through SMN can cause severe problems in many social and economic settings, such as when people rely on unverified information to make critical decisions in acute situations such as social crises (e.g., Kwon et al. 2016; Oh et al. 2013). In the runup to the 2012 and 2016 U.S. presidential elections, platforms such as Facebook and Twitter were used not only to distribute confirmed political news, but also to spread unverified and occasionally false information about candidates, which could easily be mistaken for factual information (e.g., Allcott and Gentzkow 2017; Shin et al. 2017). In a business context, online rumors and firestorms that are not addressed adequately can have severe negative consequences for companies, including the loss of trust between management, staff, and shareholders, and sustained personal and corporate reputational damage (e.g., Kimmel and Audrain-Pontevia 2010; Pfeffer et al. 2014).

Capitalizing on the topology of the social network of users, rumors can spread quickly in SMN, reaching a large audience in a relatively short time (Doerr et al. 2012). Yet both research and conventional wisdom have come to note that SMN not only can promote, but potentially also could counteract, rumor propagation: by discussing them together with others, users are said to be able to identify, challenge, and eventually correct false rumors and other forms of misinformation through an ongoing process of collective sense-making typically referred to as *self-correction* (e.g., Arif et al. 2017; Jong and Dückers 2016; Wang and Zhuang 2018). Nevertheless, evidence suggests that while there is reason to expect users will stop supporting rumors that have been proven false, they may fail to distinguish between true and false rumors beforehand (Zubiaga et al. 2016a).

In this paper, I investigate why users may fail to identify and correct false rumors although they may actually scrutinize unverified information. Explanations for this can be found in the *source and message characteristics* that can affect users'

[1] Following Kane et al. (2014), social media networks are defined as digital platforms that enable their users to set up unique user profiles, access digital content and protect it from various search mechanisms provided by the platform, establish relational connections to other users, and view and navigate these connections. This definition does not abandon the concepts of online social networks and social network(ing) sites, but rather draws a line between platforms whose functionalities rely to a significant degree upon users' relational connections—including, among others, Facebook, LinkedIn, Pinterest, Tumblr, and Twitter—and other types of social media such as weblogs, wikis, content communities, and virtual worlds.

inclination to spread or otherwise respond to a rumor (e.g., Lee et al. 2015; Li and Chong 2019; Oh et al. 2013). In addition, the *relational ties* that connect users in SMN (e.g., being friends with or following others, sending messages, or being members of the same group) can influence the flow of digital content through which a rumor is conveyed to users (Kane et al. 2014). The formation of relational ties is, in turn, contingent on the *affordances* of SMN that arise from the interactions between the platforms and their technological features, on the one hand, and user attributes and abilities on the other (Evans et al. 2017). Affordances can determine relevant actors, audiences, and publics, and thus shape the discourse that is at the very heart of self-correction (Baym and Boyd 2012).

This paper focuses on the *searchability affordance* of SMN, which allows users to access digital content through their purposeful efforts to find or discover information, rather than by navigating their relational ties to others (Boyd 2011). Using dedicated platform features such as keyword searches and automated information streams (e.g., news feeds and trending topics; Kane et al. 2014), users can follow conversations and access information shared by otherwise unrelated others, which is crucial to enable and sustain collective sense-making practices among a larger number of users (Oh et al. 2015). Searchability is also important for the propagation of false rumors, which users rarely pick up from their direct contacts (Kwon et al. 2017). Searchability can hence create opportunities for rumors to spread into distant parts of the network. This means that while searchability can enrich an ongoing discourse by integrating a wider audience, it may also interrupt or bypass collective sense-making processes. This leads to the research question: *How does searchability interfere with rumor self-correction in SMN?*

While few would likely argue that rumor dynamics are independent of the capabilities of the platforms that enable and sustain them, developing consistent and coherent explanations for how information and communications technologies (ICT) can affect social processes is a theoretical challenge. Widely used models of opinion dynamics, such as bounded confidence and voter models, do not come with inherent theories of ICT impact, which means that assumptions must be adapted from media effects and other theories—which themselves do not clearly and unambiguously account for the potential implications of SMN for networked opinion dynamics (Valkenburg et al. 2016). Affordance theory, which is well established in information systems and communications research, offers a theoretical lens for analyzing the interplay between ICT and the individual and social dynamics of human behavior, and could therefore help fill this explanatory gap. Yet to date, many of the mechanisms through which SMN might interfere with social dynamics are more or less hypothetical (Leidner et al. 2018).

This paper's goal is to identify the explanatory mechanisms through which searchability can shape the user discourse that makes sense of a rumor, and hence explain how using platform features such as keyword searches and trending topics might affect social outcomes. Agent-based simulations (ABS) are used to analyze how different modes of information access might impact collective sense-making practices in SMN. ABS are based on computer models that replicate the evolution of social systems based on the iterated and adaptive interactions of their constituent elements (i.e., *agents*; Klein et al. 2018). They hence allow for deconstructing social

dynamics that arise from the behaviors and interactions of autonomous but interdependent actors (Macy and Willer 2002)—in this case, how collective judgments regarding the veracity of a rumor arise from individual users' efforts to make sense of it. ABS thus facilitate elaborating and exploring theoretical explanations through computational experimentation (Davis et al. 2007).

Beginning from a simple theoretical rationale of how searchability might interfere with naturally occurring social influence processes, computer experiments are used to identify five theoretically and empirically plausible scenarios that characterize rumor propagation on the platform Twitter. The findings suggest that relying on dedicated content access features can enable users, under certain circumstances, to arrive at accurate rumor veracity judgments, but it can also make it difficult for them to evaluate the trustworthiness of rumor sources. Thus, searchability can help contain the propagation of false rumors, but can also prevent the spread of true rumors and corrective posts that have not (yet) been verified. The insights of this study can inform technological design and managerial governance decisions to contain the propagation and manage the adverse impacts of false rumors in SMN. In addition, they shed light on the causal mechanisms through which platform features can contribute to the emergence of social outcomes, and thus advance the theoretical integration of technological affordances and social dynamics.

In Sect. 2, I revisit major themes in prior research to facilitate a shared understanding of rumor self-correction in SMN. Based on that, I explain in Sect. 3 how the use of dedicated content access features might divert the flow of social influence among users, based on a simplified version of the rumor communication model proposed by Bloch et al. (2018). The simulation approach is outlined in Sect. 4. Subsequently, in Sect. 5, I analyze the implications of searchability for five distinct scenarios of rumor propagation on Twitter, derived from the PHEME rumor scheme dataset (Zubiaga et al. 2016b), and argue how searchability might interfere with the manifestation of self-correction in each. In Sect. 6, I discuss the paper's contributions and limitations.

2 Characterizing rumor self-correction in social media networks

According to the popular definition of DiFonzo and Bordia (2007b, p 13), rumors are "unverified and instrumentally relevant information statements in circulation that arise in contexts of ambiguity, danger, or potential threat and that help people make sense and manage risk." Analyzing, discussing, and questioning rumors together with others can help people make sense of uncertain situations and understand what is going on around them (Bordia and DiFonzo 2004). It can assist users of SMN who contribute to a conversation, as well as those following it, in keeping track of what is going on and learning about others' views, such as during crisis and disaster events (e.g., Heverin and Zach 2012; Stieglitz et al. 2018). Establishing consensus regarding the veracity of a rumor is a key element of this collective sense-making process (Bordia and DiFonzo 2004).

The term *self-correction* was first used by Starbird et al. (2014) to describe users disseminating corrections to false rumors on Twitter. Jong and Dückers (2016) refer

to self-correction more specifically as the processes through which communities of users distinguish between true and false rumors, in the course of which senders of incorrect or misinterpreted posts as well others actively try to make sense of and validate that information. Similarly, Arif et al. (2017, p 157) explain that through self-correction, crowds of users can "identify, challenge, and eventually correct misinformation." Somewhat in contrast, Wang and Zhuang (2018) argue that it is the particularly knowledgeable users from among a broader population who are enabled to identify and correct false rumors by engaging in discussions with others. Herein, the term self-correction refers to conversational practices sustained by SMN through which users resolve false rumors circulating on a platform.

Self-correction can denote different types of user behaviors. Arif et al. (2017), for instance, argue that users take actions to correct themselves as well as other users and the social information space by pointing out that they have shared false information before, deleting previously shared information, providing correct information, and addressing other users who share false information. Friggeri et al. (2014) suggest that users engage in what is called *fact-checking interventions*, posting references to fact-checking websites (e.g., Hoax-Slayer.com, FactCheck.org, PolitiFact.com, Snopes.com, and TruthOrFiction.com in the United States; Mimikama.at, HOAXmap.org, and several websites run by public service broadcasting authorities in Europe) to provide information on the veracity of a rumor.

Research, however, yields mixed evidence on user capability to distinguish between true and false rumors. Zubiaga et al. (2016a) claim that Twitter users tend to support every unverified rumor, suggesting that while they may not uphold false rumors once debunked, they are mostly incapable of distinguishing true and false rumors a priori. But while users may not be able to resolve a rumor right from the start, they may finally succeed in doing so through their concerted efforts to make sense of it. Procter et al. (2013) interpret rumor trajectories in light of a consensus-seeking process in which users gradually exchange arguments to arrive at a collective decision regarding a rumor's veracity. This is roughly in line with what Maddok et al. (2015) refer to as the *collective sense-making signature*—corrections eventually catching up with and overtaking false rumors. Carlos et al. (2013) support this claim, arguing that the overall share of tweets confirming a rumor is significantly higher if it is true, whereas there is a higher portion of posts that deny or question a false rumor.

Nevertheless, the diffusion of corrective posts was found repeatedly to lag behind that of false rumors, at least in time and occasionally also magnitude (e.g., Maddok et al. 2015; Takayasu et al. 2015). This suggests that such revisions are often less popular than the rumors they intend to correct. Inspecting rumor dynamics over time, Kwon et al. (2017) and Shin et al. (2018) find that rumor propagation can have multiple spikes, which indicates that rumors draw user attention repeatedly until they are eventually resolved. Jong and Dückers (2016) refer to this as *echo effects* — outdated and otherwise obsolete information being reshared even when an update is available from the original source.

Latecomers picking up an outdated post (Procter et al. 2013), old information being repackaged by partisan websites to look like news (Shin et al. 2018), and news media coverage that draws attention to an unresolved rumor (Spiro et al. 2012) are all potential

explanations for such effects. Zhao et al. (2016) identify a disparity between users' intention to correct false rumors and their actual rumor-combating behaviors, which means that although they might be aware of the negative consequences of false rumors and feel obliged to correct them, they may be reluctant to take action. Also, official accounts and professional journalists were found to play a major role in rumor correction (e.g., Andrews et al. 2016; Starbird et al. 2018), which casts doubts regarding the bottom-up nature of self-correction.

Further issues also arise regarding fact-checking interventions. Findings indicate that referencing fact-checking websites to debunk a false rumor does not have a significant long-term effect on the likelihood of it being reshared (Friggeri et al. 2014). Shin et al. (2017) show that the percentage of tweets rejecting a false rumor increases only slightly in response to a fact-checking intervention, and that a small percentage of users even casts doubts over these interventions' veracity. Overall, only few of the addressed users respond to such interventions in a conversation (Hannak et al. 2014) and, even if they do, responses are most often negative (Zollo et al. 2017). Also, as Wang and Zhuang (2018) suggest, most users who spread false rumors do not take corrective actions to accommodate for a debunking comment, and neither delete nor clarify rumor-supporting posts.

One possible explanation for the limited impact of fact-checking interventions is that they are rarely used to support a user, but rather to challenge others' views (Friggeri et al. 2014; Hannak et al. 2014). Furthermore, they are typically made by otherwise unrelated users (Hannak et al. 2014), which is problematic because recipients are less likely to accept objections issued by strangers rather than by their friends or followers (Margolin et al. 2018). What is more, rumor corrections cannot only fail to take effect, but may even reinforce people's misperceptions, especially if they contradict their pre-existing worldviews (Lewandowsky et al. 2012), which is known as *backfire* or *boomerang effects* (e.g., Hart and Nisbet 2012; Nyhan and Reifler 2010).

In summary, there is evidence that while self-correction can emerge from users' collective efforts to make sense of ambiguous information, their ability to identify and debunk false rumors may be limited. Studies have focused mostly on individual users' cognitive and behavioral shortcomings that might prevent them from adequately judging rumor veracity. However, they have largely considered SMN as platforms that allow users to exchange information, but do not shape their discourse. In the following section, I argue that SMN, by enabling their users to learn about a rumor not only from others with whom they share a relational connection, but also through dedicated search and query features, can help bypass social influence processes that would otherwise allow them to arrive at accurate rumor veracity judgments.

3 Searchability, social influence, and self-correction of rumors

3.1 Theoretical rationale

Rumors can spread through SMN when users not previously aware of them learn about their existence through digital content available to them, and then decide to

support the rumors themselves (Maddok et al. 2015). For instance, users can add corroborating comments to digital content that conveys a rumor, rate it positively, or reshare it, depending on a given platform's features (Lerman and Ghosh 2010). Doing so, they cannot only spread a rumor beyond the audience of an original post, but also join or follow rumor-related conversations (e.g., Boyd et al. 2010; Honeycutt and Herring 2009).

Users' stance towards digital content is not merely an expression of their independent beliefs, however, but can—at least to some extent—be influenced by the opinions of others (Li and Sakamoto 2014). *Social influence* captures the idea that an individual's beliefs, attitudes, and behaviors can be affected by others for reasons that are essentially founded in social psychology (Cialdini and Goldstein 2004). In SMN, social influence can flow when information about the activities of a user becomes available to others (Garg et al. 2011). Thus, if a user decides to support a rumor in one way or another, others can learn about that, such as when they inspect digital content posted by those they follow (i.e., their *followees*), read personal messages, or monitor contributions to a group (Kane et al. 2014). In addition, features such as status updates and notifications, as well as conversational practices like tagging, can raise user awareness of others' activities and facilitate interactions (e.g., Ellison et al. 2015; Huang et al. 2010).

Interactions along users' relational ties are frequently treated as the main mechanism of information diffusion in SMN, where the topology of ties, their contents, directionality, and the strength of information flow can shape diffusion (Garton et al. 1997). However, users' dyadic interactions typically account for only a fraction of rumor retransmissions in SMN (Kwon et al. 2017). Instead, as evidence suggests, users who spread false rumors often learn about them through dedicated content access features (e.g., Carlos et al. 2013; Gupta et al. 2013; Vosoughi et al. 2018). Searchability might hence help spread a rumor beyond its original audience, based on platform features such as keyword searches and trending topics that make it easy for users to locate digital content beyond their direct contacts (Boyd 2011). Thus, searchability could divert the flow of social influence among users, which means it could affect collective sense-making by manipulating the structure of interactions and what information is available to actors when making up their minds (Mason et al. 2007).

Another question is whether searchability could imply qualitative changes in the operating logic of social influence. Expanding on Garg et al. (2011), one might argue that users who discover information using dedicated content access features do not observe others' behaviors, but rather see decontextualized content that does not transmit social influence. However, in contrast to web searches, for instance, information obtained through the dedicated content access features of SMN cannot be expected to come detached from its social context (e.g., Elsweiler and Harvey 2015; Teevan et al. 2011). For one, many of these features rely on users' connections to identify potentially relevant content (Kane et al. 2014). Furthermore, it is not necessary for two actors to be related directly to influence each other, as mechanisms such as structural equivalence (e.g., Zhang et al. 2018) and information sampling (e.g., Denrell and Le Mens 2017) can result in their adapting to a shared social environment.

From a theoretical standpoint, this issue is about the distinction between *peer influence* (i.e., behavioral adaptation motivated by social conformity) and *social learning* (i.e., behavioral adaptation based on rational deliberation). Both model the adjustment of actors' attitudes or behaviors in response to information available to them from others. The main difference is that models of peer influence tend to focus on the social processes through which actors' relationships determine information exchanges, whereas social learning approaches typically emphasize the decision-making rules through which actors revise their beliefs based on different sources of public and private information. Thus, the question is not whether actors' judgments can be affected by those of others, but rather how such opinion changes and the information behaviors that result from those changes might ultimately be motivated.

I hence argue that searchability can alter the path through which social influence, in a broad sense, is transmitted among SMN users: dedicated content access features create the conditions for users to learn about a rumor before they might otherwise have from their direct or indirect contacts, or find out about responses that differ from the views expressed within their immediate social environment. On the one hand, this could increase the variety of opinions and allow a larger number of users to participate in collective sense-making, which in turn could enhance the quality of discourse and support the spread of corrections to false rumors. On the other hand, however, false rumors already debunked in one part of the network could be transferred into others, and an unsolicited influx of arguments could prolongate collective sense-making. Searchability could, therefore, not only promote the diffusion and persistence of false rumors, but also help maintain high levels of uncertainty, which is problematic if it prevents users from taking appropriate actions.

I formalize these assumptions below, based on the rumor communication model of Bloch et al. (2018). This model is the basis for using ABS to identify the mechanisms through which changes in the structure of information flow might counteract conversational practices that are typically referred to as self-correction.

3.2 Model formulation

Most research has relied on social contagion models to analyze rumor dynamics in SMN (Serrano et al. 2015). These models' assumptions, however, stand in stark contrast to the notion of social influence delineated above, relying on mere contact rates between users to account for diffusion rather than on social forces that drive adaptation (Young 2009). Alternative models of networked opinion dynamics make claims that are more sophisticated in this regard; still, they lack assumptions about people's rumor-spreading behaviors and motives, and they rarely account for the impacts of ICT beyond maybe that of mass media (Sîrbu et al. 2017).

This paper hence relies on a simplified version of the rumor communication model suggested by Bloch et al. (2018), who describe rumor propagation in social networks and public broadcast environments based on actor decisions to create and forward rumor messages according to their immanent behavioral intentions and the messages' presumed veracity. SMN are incorporated as a hybrid of networked and public broadcast environments in which users' relational ties

introduce instances of networked communication that complement information access through dedicated content access features.

The model is based on a population of $N = \{1, \ldots, n\}$ agents inhabiting a world with two possible states of nature $\theta \in \{0, 1\}$. Let $x_i(t)$ denote the ith agent's belief that $\theta = 1$ at time $t \in \{0, 1, 2, \ldots\}$, where agents share a prior belief $x(t = 0) = \pi \forall i \in N, 0 < \pi < 0.5$. In the context of this paper, agents represent the individual users of SMN, and the state of nature θ corresponds to the veracity of a rumor. To facilitate the subsequent explanations, let $\theta = 1$ denote that a rumor is true. At $t = 0$, one agent privately receives a perfect signal s $\in \{0, 1\}$ that corresponds to the true state of nature. Each agent is equally likely to receive this signal. The recipient agent i then creates a message $m_i(t = 0|s) \in \{0, 1\}$ that passes the rumor to other agents within the population.

Each agent obeys the rules of one of two communication strategies. *Unbiased agents* share only information they believe to be true. If an unbiased agent receives the initial signal, they pass it on without modification, as its veracity is undisputable, which implies that $m_i^u(t = 0|s) = s$. *Biased agents*, in contrast, have a special interest in spreading the word that a rumor is true. Thus, if a biased agent receives the initial signal, that agent always creates the message $m_i^b(t = 0|s) = 1$. In the context of SMN, users might, for instance, want to provide others with timely and presumably relevant information about an uncertain event, such as a natural disaster (e.g., Abdullah et al. 2017; Li et al. 2014). Biased agents' behavior can thus be interpreted as a function of their individual motives rather than as attempts to achieve some sort of collective outcome, and without implying a normative judgment.

Once a message has been created, it is disseminated to other agents within the population, who then decide whether to propagate it further. Rumor transmission takes place in discrete time, where at each time step $t + 1$, every agent j who has received a message $m_i(t)$ from a preceding agent i, $i, j \in N$, $i \neq j$, decides whether to retransmit it. Agents can decide either to relay or block a message, which is denoted as $m_j\big(t + 1|m_i(t)\big) \in \{m_i(t), \emptyset\}$. Agents, though, cannot alter the message contents. Thus, once a message has been created that does not correspond to the true state of nature (i.e., a *false rumor*), it remains in circulation, although further messages might be created subsequently to correct it.

Again, biased agents only spread the message that a rumor is true, regardless of their veracity beliefs. This results in the communication strategy:

$$m_j^b\big(t + 1|m_i(t)\big) = \begin{cases} m_i(t) & \text{if } m_i(t) = 1 \\ \emptyset & \text{else} \end{cases} \tag{1}$$

Unbiased agents spread a message only if they believe it to be true. Otherwise, they do not retransmit the message. Upon receiving a message $m_i(t) = 0$, unbiased agents always retransmit the rumor, as it could only have been created and passed on by unbiased agents, which means that it definitely corresponds to the true state of nature. If, however, they receive a message $m_i(t) = 1$, they will relay it only if the likelihood that it was created by a biased agent is sufficiently low. This corresponds to the communication rule:

$$m_j^u(t+1|m_i(t)) = \begin{cases} m_i(t) & \text{if } m_i(t) = 0 \\ m_i(t) & \text{if } m_i(t) = 1 \text{ and } x_j(t+1) > 0.5 \\ \emptyset & \text{else} \end{cases} \quad (2)$$

Unbiased agents are crucial to keeping diffusion alive: they can disrupt propagation, but if they continue to spread the message that a rumor is true, it abets the efforts of biased agents who themselves wish to propagate that message. Thus, self-correction essentially depends on unbiased agents' ability to evaluate the veracity of a rumor message effectively.

Agents apply Bayes' rule to revise their rumor veracity beliefs according to $x_j(t+1) = \pi/(b_j(t^*) + (1 - b_j(t^*)) \cdot \pi)$, where $b_j(t^*)$ denotes the percentage of biased agents among those who have at some prior $t^* \in \{0, \ldots, t\}$ participated in the rumor transmission process. Following Bloch et al. (2018), rumor propagation can take place in networked and public broadcast environments. In both settings, the identity of the agent who has received the initial signal is unknown and must be inferred by subsequent agents. In the *public broadcast environment*, agents' messages can be transferred to one or more other agents simultaneously. This corresponds to situations in which information diffuses through broad-based media, such as websites, newspapers, or dedicated content access features in the context of SMN. In this case, only the message itself is transferred, and the communication strategies of agents who (re)transmit it are private knowledge. Assuming that the overall number of biased agents within the population $B \subset N$ is commonly known, the probability that a message was originally created by a biased agent is $b_j(t^*) = |B|/(|N| - 1)$ for any unbiased agent j.

In the *networked communication environment*, agents communicate pairwise (e.g., in personal conversations, via e-mail, or through dyadic interactions in SMN), and their communication strategies are commonly known. This implies that all pathways through which a rumor could potentially have been transmitted to a recipient agent j can be retraced, except again for the original source of information. In other words, agents are assumed to keep track of who has been involved in the transmission of a rumor message, although they may not pay heed to the exact transmission sequence. In the context of SMN, this corresponds to the ability of users to view and traverse their relational connections, as well as those made by others on the platform (Kane et al. 2014). Let $G(N, t^*)$ represent the directed social network that results from agents' dyadic interactions along their relational ties, where $g_{ij}(N, t^*) \in \{0, 1\}$ represents the communication link between two agents i and j, $i, j \in N$, $i \neq j$, and where $g_{ij}(N, t^*) = 1$ if agent i has previously transmitted at least one message to agent j, and $g_{ij}(N, t^*) = 0$ else.

For any unbiased agent j who has received the message $m_i(t) = 1$ from one of their neighbors i, $b_j(t^*)$ depends on the communication strategies of the agents along the communication pathway who might have received the initial signal. If, for instance, j has received the message that a rumor is true from an unbiased neighbor i who in turn has previously interacted with a biased agent k, the signal could originally have been created either by j (in which case it is definitely true) or by k and then transmitted to j (in which case k could have misrepresented the signal in the first place).

The subset of potential rumor origins for a focal agent j who has received a message $m_i(t)$ is denoted as $N_j(i,t^*) = \left\{k | g_{ij} \in G(N,t^*) \wedge g_{ki} \in G(N,t^*)^{[R]}\right\}$, $i,j,k \in N$, $j \neq i,k$, which describes the set of all agents k who are directly or indirectly connected to j through i, including i themselves (the reachability matrix $G^{[R]} = \{g_{ki}^{[R]}\}$ denotes the direct and indirect connections between any two agents in the network, where $g_{ki}^{[R]} = 1$ if a directed path of arbitrary length exists from k to i, and $g_{ki}^{[R]} = 0$ else; Wasserman and Faust 1994). This results in $b_j(t^*) = |B_j(i,t^*)|/|N_j(i,t^*)|$ for the networked communication environment, with $B_j(i,t^*) \subset N_j(i,t^*)$ denoting the subset of biased agents among those who could have received the initial signal of a message that was transmitted to j through i at time t. Agents do not discount messages received repeatedly from potentially the same source, which can increase the relative influence of well-connected agents (DeMarzo et al. 2003).

Bloch et al. (2018) demonstrate that unbiased actors can effectively prevent the spreading of false rumors if the likelihood that a biased agent created the message is sufficiently low. In the public broadcast environment, agents evaluate the veracity of a rumor based on the overall share of such agents across the population, whereas in a networked setting, retracing the potential transmission paths of a rumor enables them to block rumors from parts of the network that are dominated by biased agents. Therefore, both communication environments provide mechanisms that allow truth-seeking agents to hinder the propagation of false rumors.

This paper introduces *hybrid communication environments*, which relax the assumption that agents can keep track of *all* agents who have been involved in rumor communication while granting them the ability to follow conversations among their direct and indirect neighbors. In a hybrid communication environment, an agent j who receives a message from a directly related agent i can retrace the potential diffusion pathways over all agents who are directly or indirectly connected to i, just as in the networked setting. If there are however further unrelated agents, it is possible that *any one* of them could have received the initial signal, which would then have been taken up by an agent from within their connected component. Hence, in addition to the set of directly and indirectly connected agents, j would also have to consider the likelihood of an unconnected information source being biased to evaluate the veracity of a message, which is denoted as:

$$b_j(t^*) = \frac{|B_j(i,t^*)| + \left(|B| - |B_j(i,t^*)|\right)/\left(|N| - |N_j(i,t^*)| - \mathbf{1}\right)}{|N_j(i,t^*)| + \mathbf{1}_{\mathbb{N}_0}\left(|N| - |N_j(i,t^*)| - \mathbf{1}\right)} \tag{3}$$

The indicator function in the denominator is equal to one if $|N| - |N_j(i,t^*)| > 1$, which is, if $N_j(i,t^*)$ is a proper subset of N. Equation 3 thus denotes the overall likelihood that the original source of a rumor message transmitted to an unbiased agent j in a hybrid communication environment at a time t is biased. It combines the share of biased agents within the focal agent's connected component and the probability that an agent in the remainder of the population is biased, assuming that either one of j's direct or indirect neighbors or one other agent could have received the initial signal. If the share of biased agents differs substantially between the focal agent's connected component and in the remainder of the population, this could interfere with unbiased agents' ability to evaluate the trustworthiness of transmitters. Hybrid

communication structures that allow agents to learn about a rumor both through interactions between personally known others and from public information sources that obscure message creators' credibility could hence disrupt the chain of accountability between related agents, and shift the ratio of biased and unbiased agents in the inferred transmission pathway. Thus, searchability might divert the flow of social influence in a way that counteracts self-correction.

The model does not explain, however, *how* searchability might affect self-correction. This is where ABS come in: they serve as analytical tools to analyze the communication structures that result from different modes of information access, and evaluate their potential implications for the flow of social influence. The methodological approach is described in the section that follows.

4 Methodology

4.1 Research approach

The purpose of this paper is to develop a theoretical explanation of how the searchability affordance of SMN—that is, users relying on dedicated content access features, rather than their relational ties to other users, to learn about and make sense of a rumor—might divert the flow of social influence and hence impede self-correction. Thus far, I have proposed a formal model for how searchability can influence the decisions of actors who seek to support true rumors while blocking the forwarding of false ones (i.e., of *unbiased agents*). This section discusses the use of ABS to analyze this model and identify parameter ranges that are both theoretically plausible and realistic in that they can reproduce the macro-level characteristics of empirically observed rumor conversations on Twitter.

ABS rely on computer simulations to analyze social dynamics based on the behaviors and interactions of autonomous, yet interdependent, actors (i.e., the *agents* that constitute a social system) that follow simple behavioral rules and adapt to how they experience their environment (Macy and Willer 2002). The key benefit of ABS is that they allow for describing and analyzing emergent macro-level behaviors based on the micro-level behaviors of the constituent agents, even when the mathematical descriptions of these behaviors are not analytically tractable (Klein et al. 2018). Thus, ABS can facilitate insights into the generative processes that lead to emergent social behaviors, including nonlinear, conditional, and qualitative effects (Smith and Conrey 2007).

Computational model representations can run the gamut from highly abstract *toy models* to empirically rich descriptions of real-world social systems (Klein et al. 2018). The present work is a *typification* that investigates the properties of a broader class of diffusion phenomena (Boero and Squazzoni 2005), and is hence situated in between: while analyses rely on the PHEME rumor scheme dataset (Zubiaga et al. 2016b) to place analyses in a real-world context, ABS are used to reveal the causal mechanisms that might underlie self-correction, rather than describe or predict how false rumors diffuse or persist in SMN.

Below, I describe the computer model's main constructs, parameters, and procedures and the steps that were taken to validate and calibrate the model.

4.2 Computational representation

Rumor messages created by the members of a denumerable agent population constitute the basic construct of the simulation model. Messages are distinguished according to the information they convey (i.e., whether they state that a rumor is true or false) and their veracity (i.e., whether this information corresponds to the true state of nature). Rumor transmission is modeled based on the pathway through which a message spreads from an external source of information along a *sequence of agents* who retransmit it. Individual agents are characterized by their communication strategy, as described in Sect. 3.2: biased agents retransmit messages according to which a rumor is true, regardless of its veracity; unbiased agents evaluate the veracity and retransmit only if they believe the rumor to be true. The likelihood of an agent being biased p_{bias} is exogenously specified.

The model's main boundary condition is the *mode of information access*. Agents have two options to receive rumor messages: they can learn about a rumor from other agents with whom they share a relational connection, or they can gain access to the messages created by *any* other agent through dedicated information access mechanisms. Whether unbiased agents arrive at an adequate judgment of a rumor's veracity thus depends on (a) their prior beliefs regarding rumor veracity, (b) the relative share of biased agents within their connected component and in the remainder of the population, and (c) the agents' positions within the retransmission sequence.

This setup implies that manifested information exchanges between agents are analyzed, rather than the latent structure of *relational ties* that enable them. The relational ties between any two agents in the network are characterized by an exogenously specified likelihood p_{tie}, $0 \leq p_{tie} \leq 1$, with which they are activated, where $p_{tie} = 0$ corresponds to the ideal-type public broadcast environment and $p_{tie} = 1$ represents the networked communication environment. In addition, homophily h (i.e., the extent to which agents prefer connecting with similar others in terms of the communication strategy pursued; McPherson et al. 2001), $0 \leq h \leq 1$, determines whether ties are activated, conditional on the agents' communication strategies, where $h = 0$ implies that agents activate relational ties to biased and unbiased agents with the same probability and $h = 1$ indicates that they only connect to similar agents, provided there are any such agents. In the context of the model, homophily can help unbiased agents prevent the spread of false rumors by blocking messages from parts of the network dominated by biased agents (Bloch et al. 2018).

At the setup of each simulation run, one agent is created, who then learns about a rumor from an exogenous source (i.e., the computer program) and initializes a rumor message according to that agent's communication strategy. In each subsequent iteration, one further agent is created and receives a rumor message from a previous agent, either by activating a relational tie to one randomly selected prior agent or by accessing the message created by the directly preceding agent, contingent on p_{tie} and h. Upon receiving a message, agents revise their beliefs and create their own

messages in accordance with their communication strategies. Biased agents who receive the message that a rumor is false, as well as unbiased agents who arrive at the belief that the message does not correspond to the truth, do not create messages, but exit the simulation. A simulation run ends when unbiased agents' beliefs are stable and no longer subject to changes in further iterations. Rumor self-correction is said to be effective if, as a result of the retransmission process, an arbitrary unbiased agent would arrive at a veracity judgment that corresponds to the true state of nature, that is, if $\lim_{t\to\infty} x_i = \theta$. To evaluate this criterion, an additional unbiased agent is created after each iteration; those agents updates their beliefs according to the same rules as regular agents, but exit the model immediately after updating and thus have no further implications for rumor propagation.

In the following section, I describe the structural validation steps taken to ensure that the computer model truthfully represents the theoretical model described in Sect. 3.2.

4.3 Structural validation

The goal of structural validation is to ensure that the computer model resembles the assumed propagation processes as closely as possible (Manson 2003). The model was implemented in *NetLogo* (version 6.0.2; Wilensky 1999). Experiments were implemented and run using *R* (version 3.6.1; R Core Team 2019) and the package *nlrx* (Salecker et al. 2019). For *face validation*, the computational routines were continually traced to ensure they produced the intended behaviors. Subsequently, for *sensitivity analysis*, a wide range of input parameter configurations were tested to compare simulation outcomes with the analytical results (Kleijnen 1995), which are available from Bloch et al. (2018) for the ideal-type networked and public communication environments.

Following the recommendations of Marino et al. (2008) and Thiele et al. (2014), Latin Hypercube Sampling was used to generate $N = 5000$ model input parameter configurations for both communication environments, namely, for agents' prior beliefs, the likelihood of their being biased, and for the extent of homophily (in the networked communication environment only). As an outcome variable, the veracity judgments at which an arbitrary unbiased agent would arrive at the end of a simulation run (i.e., whether they believed a rumor to be true or false) were measured. Logistic regressions were fitted to determine the strength of impact of each of the model input parameters on unbiased agents' emergent veracity beliefs. Table 1 shows the parameter estimates.

Overall, unbiased agents arrived at veracity judgments that corresponded to the rumors' actual veracity in 43.8% of all simulation runs. Numbers differ significantly across the two communication environments: In the public broadcast environment, agents arrived at correct judgments in 47.6% of the experiments, whereas in the networked communication environment, they did so in only 40.1% ($p < 0.01$). In total, rumor self-correction was hence effective in a little less than half of all cases. This is not a pessimistic outcome, seeing that unbiased agents are naïve in the sense that

Table 1 Logistic regression coefficients of unbiased agents' emergent rumor veracity beliefs in the public broadcast and networked communication environment ($N = 5000$ each)

	Public broadcast environment $\widehat{\beta}$	Networked communication environment $\widehat{\beta}$
Intercept	− 0.862***	0.752**
p_{bias}	− 13.799***	− 9.007***
π	14.747***	7.144***
h	–	− 0.146
p_{bias}:π	16.118***	4.278**
p_{bias}:h	–	− 0.870
π:h	–	2.162
McFadden's R^2	0.598	0.478

***$p < 0.01$; **$p < 0.05$; *$p < 0.1$

their prior beliefs are not correlated to rumor veracity and they are unable to learn (e.g., to improve their prior beliefs incrementally or to restrict communications to a subset of evidently trustworthy agents).

Inspecting the logistic regression coefficients as well as the distribution of emergent beliefs in Fig. 1 supports the conclusion that unbiased agents in both communication environments are more likely to believe a rumor to be true if the percentage of biased agents is low, and if furthermore their prior beliefs are in favor that judgment. This pattern is robust when controlling for rumor veracity. A lower share of biased agents does not in all cases, therefore, promote the emergence of correct veracity judgments, as it can also mislead unbiased agents to spread untruthful messages. The outcomes are approximately in line with the full communication equilibrium identified by Bloch et al. (2018), according to which unbiased agents would spread the message that a rumor is true iff $b_j(t^*) \leq \pi/(1-\pi)$.

Homophily does not have a significant impact on unbiased agents' veracity judgments in the networked communication environment, although it should enable them to evaluate the reliability of potential rumor origins more accurately. However, it is possible that the relationship between rumor veracity and unbiased agents' judgments is not monotonic: homophily might not only make it easier for them to block false messages, but also to put faith in true ones, which means that positive and negative effects of homophily might cancel each other out. The implications of homophily are discussed in more detail in Sect. 5.

Below, I describe how the computer model was then calibrated, using the PHEME rumor scheme dataset (Zubiaga et al. 2016b) to match it with empirical patterns of rumor propagation on Twitter.

4.4 Empirical calibration

Analyses rely specifically on those parameter configurations for which the computational model can reproduce the patterns of empirically observed rumor conversations. The micro-level model constructs are not empirically observable; therefore, the *indirect calibration* approach described by Windrum et al. (2007) was used to

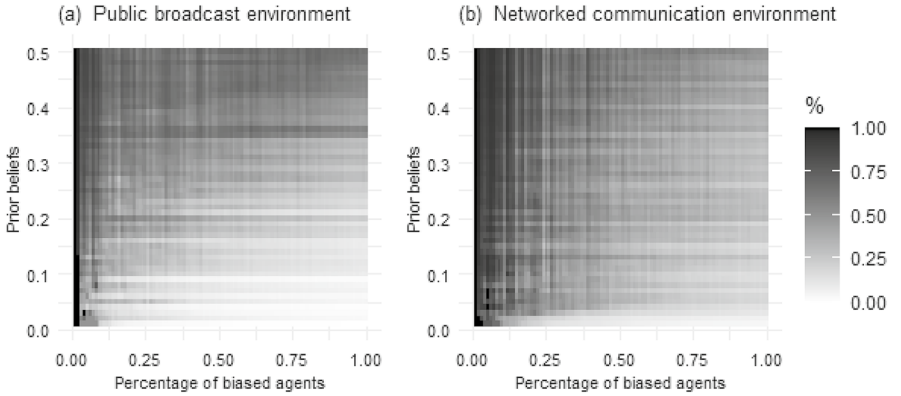

Fig. 1 Percentage of experiments in which unbiased agents judged a rumor to be true, public broadcast (**a**) vs. networked communication environment (**b**)

identify input parameter ranges that produce outcomes in line with two macro-level stylized facts, namely, the average number of messages that support (rather than deny or comment on) a rumor, and the relative likelihood with which users participate in a conversation in response to a message from a user with whom they share a relational connection (rather than unrelated others). Overall, five clusters of parameter configurations were identified that are consistent with the observed macro-level patterns, which are then used to develop explanations about the implications of network connectivity for the self-correction of rumors on the platform Twitter.

The model was calibrated using the PHEME rumor scheme dataset, which contains tweets and annotations for the propagation of rumors associated with nine breaking news events on Twitter (Zubiaga et al. 2016b). While Twitter is generally oriented toward digital content rather than the connections between users, it has adopted networking features characteristic of SMN (Berger et al. 2014), and is therefore typically treated as an instance of SMN (e.g., Kane et al. 2014; Karahanna et al. 2018). Details on data collection and annotation are available from Zubiaga et al. (2015, 2016a).

Analyses rely on a subset of tweets in English pertaining to rumors regarding five breaking news events (as opposed to longstanding rumors): the unrest in Ferguson, Missouri in 2014; shootings on Parliament Hill in Ottawa, Canada in 2014; the Lindt Cafe siege in Sydney, Australia in 2014; the Charlie Hebdo shooting in Paris in 2015; and the crash of a Germanwings passenger aircraft in 2015. The basic units of analysis are the *conversations* pertaining to each event, based on which it is possible to analyze how users relate to each other in their rumor responses (Kogan and Palen 2018). Each conversation consists of a source tweet that started a rumor and one or more tweets that responded using Twitter's *@reply* feature. After removing conversations in which the source tweet did not take a clear stance on a rumor (and in which it was therefore not possible to determine whether subsequent messages agreed with it), a total of 266 conversations and 4744 tweets remained for empirical calibration.

Empirical calibration is based on annotated information that classifies tweets with respect to whether their authors agreed with a rumor and whether they responded to a tweet of a user with whom they were connected through a unidirectional follower relationship. The latter corresponds to the density of the information flow network, assuming that users learned about a rumor from their followees who have responded to it previously (Kwon et al. 2017). Table 2 is an overview of the macro-level attributes of the conversational threads.

The software tool *BehaviorSearch* (version 1.10; Stonedahl and Wilensky 2010) was used for input validation—that is, to explore the parameter space and identify parameter configurations that produce the desired macro-level outcomes (Fagiolo et al. 2019)—using Genetic Algorithms. The model is underdetermined, as multiple combinations of the same micro-level input parameters can produce each macro-level outcome. For instance, unbiased agents might judge a rumor to be false because their prior beliefs regarding rumor veracity are low or because most agents who could have created the message are biased. Similarly, they might activate relational ties to unbiased rather than biased agents because the level of homophily is high or because there are only a few biased agents overall to whom they might connect. To identify as many plausible input parameter configurations as possible, parameter identification was repeated ten times for each of the conversations, treating the outcomes of each cycle as independent inputs for the subsequent analyses.

Many parameter configurations that reproduce the empirically observed macro-level outcomes are similar, which implies that there might be distinct patterns that underlie rumor propagation. Five consistent clusters were detected using hierarchical agglomerative clustering with Ward's minimum variance method (cophenetic correlation $c = 0.710$). The characteristic structures that result from the parameter configurations within each cluster are hereafter referred to as *scenarios*. Table 3 is an overview of the clusters identified.

Each cluster corresponds to a set of micro-level input parameter ranges that reproduce approximately the observed macro-level patterns of rumor propagation. They should not be understood as descriptions of the real-world social system. Instead, the scenarios are used to characterize the structures of information flow that result from different configurations of network connectivity and explain how these different patterns might be related to the manifestation of self-correction. While the identified parameter configurations could serve as inputs for further experimentation

Table 2 Overview of the macro-level attributes of rumor conversations associated with five breaking news events

Breaking news event	#Threads	#Tweets	Density			Rumor support		
			Min	Avg	Max	Min	Avg	Max
Ferguson unrest	41	1159	0.000	0.350	0.800	0.051	0.191	0.529
Ottawa shooting	58	844	0.000	0.316	1.000	0.000	0.273	0.833
Sydney siege	71	1236	0.000	0.426	1.000	0.029	0.246	0.750
Charlie Hebdo shooting	72	1204	0.000	0.547	1.000	0.000	0.281	0.895
Germanwings plane crash	24	301	0.000	0.373	0.875	0.047	0.326	0.750

Table 3 Input parameter ranges identified through empirical validation

p_{biased}			π			p_{tie}			h		
Min	Avg	Max	Min	Avg	Max	Min	Avg	Max	Min	Avg	Max
Cluster 1 (24.1%)											
0.730	0.894	0.990	0.010	0.054	0.180	0.000	0.079	0.190	0.820	0.963	1.000
Cluster 2 (21.5%)											
0.440	0.793	0.940	0.010	0.047	0.170	0.000	0.031	0.140	0.190	0.396	0.670
Cluster 3 (11.3%)											
0.190	0.301	0.370	0.010	0.048	0.070	0.000	0.030	0.090	0.300	0.489	0.890
Cluster 4 (23.1%)											
0.860	0.885	0.920	0.120	0.197	0.390	0.000	0.016	0.070	0.340	0.659	0.750
Cluster 5 (20.0%)											
0.840	0.921	0.970	0.030	0.099	0.120	0.190	0.227	0.450	0.400	0.451	0.490

and analyses (Melamed et al. 2012), this would merely allow for quantifying effect strengths within the narrow boundaries of the theoretical model. This paper, conversely, develops a theoretical rationale for the implications of different modes of information access on collective sense-making outcomes.

In the following section, I discuss the characteristic structures of information flow that result from the input parameter configurations for each scenario. Based on that, the potential implications of searchability for unbiased agents' emergent veracity judgments are analyzed.

5 Implications of searchability for the self-correction of rumors

5.1 Scenario overview

The two input parameters that have the strongest impact on unbiased agents' emergent rumor veracity beliefs in the ideal-type public and networked communication environments are the share of biased agents (i.e., of agents who would spread rumor messages regardless of their veracity) and agents' prior beliefs (i.e., their a-priori expectations that a rumor is true). The input parameters for the *share of biased agents* are higher than 75%, on average, in all scenarios except scenario 3, which is characterized by a moderate share of biased agents of about 30%, on average. As regards agents' *prior beliefs*, in all scenarios except scenario 4, prior beliefs of less than 10%, on average, produce outcomes that resemble the empirically observed conversation patterns; in scenario 4, unbiased agents assign an a-priori likelihood of about 20% to a rumor being true, on average.

The subsequent explanations focus on how different patterns of information flow network connectivity might affect self-correction. Connectivity depends on two model input parameters. The first is *network density*, which describes the likelihood with which agents activate their relational ties to others to learn about a rumor; it is below

10% in scenarios 1 through 4, on average, whereas scenario 5 is characterized by an average density of about 23%. The second is *homophily*, for which there is considerable variation across the scenarios. Three scenarios—scenarios 2, 3, and 5—are characterized by moderate extents of homophily of between 40 and 50%, on average, which implies that unbiased agents are about twice as likely to receive information from other unbiased agents if they activate their relational ties to others rather than obtaining information from an unconnected member of the agent population. In scenario 4, average input values for homophily are about 65%, and in scenario 1 about 96%.

Below, I discuss the patterns of network connectivity that result from the respective parameter configurations, and how they might affect unbiased agents' ability to distinguish between true and false rumors.

5.2 Implications of searchability in each scenario

5.2.1 Insights from the scenarios 1 and 2

The first scenario, which comprises about 24% of the input parameter configurations identified, is characterized by low values of network density, low prior beliefs, and high shares of biased agents. The extent to which homophily dictates tie formation is high. Unbiased agents are hence, on average, more than 90% more likely to learn about a rumor from a trustworthy agent than from a random member of the agent population if they rely on their relational ties to access information. The network structures that result from these parameter values resemble the ideal-type public broadcast environment, as agents receive information from their followees only sporadically. It is hence unlikely that unbiased agents would support a rumor regardless of its veracity, as the likelihood that it was initially released by a biased agent is unduly high. This implies that while false rumors would be effectively blocked, unbiased agents would support neither true rumors nor corrective posts.

However, occasional instances of networked communication could help unbiased agents distinguish between true and false rumors, as the likelihood that they will receive trustworthy information through their relational connections is considerably higher than when accessing information from unrelated others. Buskens (1998), for instance, finds that agents' ability to communicate their trustworthiness (i.e., their communication strategies) to those with whom they are connected is positively related to the level of trust in the network, which in turn can increase the quality of information transferred between actors (Droege et al. 2003). Knowing whether those who could have created a message would be willing to spread misinformation could thus allow agents to place trust more selectively.

As for self-correction, this insight implies that social relationships that allow agents to receive information—preferably from trustworthy others—might enable them to recognize true rumors more easily and selectively support those messages. Dyadic exchanges might therefore create trustworthy transmission pathways for individual messages. This mechanism depends not only on the relative frequency of networked communication, but also on whether agents can depend on their relational ties to provide them with information from trustworthy others. While lower

Fig. 2 Rumor transmission along agents' relational ties creates trustworthy transmission pathways for individual messages, even when most agents are biased

values of network density may limit the rates at which this mechanism is activated, it might still be effective if tie formation is subject to high levels of homophily.

Figure 2 illustrates this mechanism. The example consists of $n = 10$ nodes who participate in a rumor conversation, 6 of which are biased and 3 unbiased. Undirected edges represent the relational connections along which agents can retrace communications. In an ideal-type public broadcast environment, the share of biased agents who could have received the initial signal would be $b = 0.667$ for any unbiased agent, which means they would not spread the message that a rumor is true. However, occasional instances of networked communications allow unbiased agents to distinguish between messages originating from different parts of the network: if, for instance, $U1$ learns about a rumor from neighbor $U2$, the likelihood that it was created by a biased agent (i.e., by $B1$ or one of the unconnected agents) drops to $b_{U1} = 0.458$. If all agents in the population were directly or indirectly related to the focal agent, it would allow them to evaluate rumor veracity even more accurately, as rumors originating in parts of the network dominated by biased agents could be blocked (Bloch et al. 2018). In the hybrid case, however, dyadic exchanges can open up pathways for individual messages to be transmitted through a sequence of largely trustworthy agents.

The second scenario accounts for about 22% of the input parameter configurations identified and is similar to the first, except for its lower values of homophily of about 40%, on average. Just as in the first scenario, relying on dedicated content access features prevents unbiased agents from spreading a rumor. However, lower values of homophily imply that the way unbiased agents learn about a rumor does not make a large difference. This indicates a positive relationship between the extent to which tie formation is subject to homophily and whether it provides agents with meaningful information about the trustworthiness of a rumor source: for lower values of homophily, the likelihood of receiving trustworthy information either from a direct contact or from unrelated others converges. Therefore, even if unbiased agents receive information from their neighbors, they may not be able to determine whether it was transmitted through an uninterrupted chain of trustworthy others. While this would still allow them to impede the propagation of false rumors, as they would simply assume all messages to be false, it would not enable them to promote true rumors and corrective posts.

5.2.2 Insights from the Scenarios 3 and 4

The third scenario accounts for about 11% of the input parameter configurations identified and is similar to the first two, except that the share of biased agents is lower as, on average, only about 30% of the agent population is biased. The third

Fig. 3 Rumor transmission along agents' relational ties rebalances the shares of biased and unbiased agents

scenario is hence the only one in which the majority of agents is unbiased, which implies that there might be cases covered by the full communication equilibrium (i.e., in which unbiased agents would spread the message that a rumor is true). Unbiased agents' prior beliefs are still low; on average, they expect a rumor to be true with an a-priori likelihood of about 5%. In the ideal-type public broadcast environment, full communication would hence be possible for a sufficiently low share of biased agents. However, depending on the composition of agents within their connected component, instances of networked communication could lower the threshold value for the share of biased agents for which unbiased agents would continue to spread rumor messages.

Figure 3 is an example of how social networks might motivate unbiased agents to spread rumor messages. The population consists again of $n = 10$ agents, 5 of which are biased and 5 unbiased each. The overall likelihood that a rumor message was created of $b = 0.556$ for any unbiased agent in the public broadcast environment. If, however, a focal agent $U1$ can rule out the possibility that a rumor was created at least by *some* biased agents whose communications that focal agent can monitor, it could increase the trust messages received from the remainder of the network. In the example, two edges suffice to decrease the likelihood of a biased message source for a message received an *unrelated* agent to $b_{U1} = 0.429$. Agents whose prior beliefs are sufficiently in favor of a rumor could thus be convinced to keep spreading messages.

In contrast to the first two scenarios, this mechanism is not so much about spreading individual messages through a sequence of trustworthy agents, but rather about creating a communication environment in which unbiased agents are generally willing to spread a rumor, as occasional instances of networked communication manipulate the relative share of unbiased agents within their connected component and the population as a whole in favor of messages' presumed veracity.

The fourth scenario accounts for about 23% of the input parameter configurations identified; its parameters are again similar to the first two scenarios. What distinguishes this scenario is that unbiased agents' prior beliefs are more moderate as, on average, they expect a rumor to be true in about 20% of the cases. Therefore, the unbiased agents' prior beliefs provide the opportunity to move unbiased agents towards the full communication equilibrium by increasing the threshold value of biased agents for which they would support a rumor, along the lines of the explanatory mechanism hypothesized for the third scenario.

5.2.3 Insights from scenario 5

The fifth and final scenario accounts for about 20% of the parameter configurations identified; it is distinct from the others as the overall likelihood with which

agents activate their relational ties to access information is higher. In about 23% of the cases, on average, they rely on their social relationships rather than dedicated features for accessing digital content. The other input parameters are similar to the other scenarios: there is a high percentage of biased agents, and unbiased agents' prior beliefs that a rumor might be true are low. Tie formation is subject to moderate degrees of homophily.

The patterns that could emerge from users' relational ties in this scenario are interesting because network theories typically discuss the implications of macro-level structures of connectivity, rather than those of occasional dyadic connections. The key argument is that network closure can facilitate social action, as it supports trustworthy interactions based on shared social norms and expectations (e.g., Coleman 1988; Granovetter 1985). While the identified ranges of network connectivity might not be sufficient to ensure consistently high levels of exchange, they are not overly low when compared to empirically observed sparse social networks in which communities of actors are rarely connected to the remainder of the network (Leskovec et al. 2008).

Considering the potential implications of homophily, agents are again about 45% more likely to receive messages from trustworthy agents through their relational ties. Interactions with similar others might therefore allow unbiased agents to spread presumably true rumors among communities of unbiased agents, although false rumors might continue to propagate among biased agents. Figure 4 illustrates this mechanism, based on a network with $n = 10$ agents, 6 of whom are biased and 4 unbiased. Only agents of the same type are connected through relational ties. In an ideal-type public broadcast environment with the same agent composition, the likelihood that a message was created by a biased agent is $b = 0.667$ for all unbiased agents. However, for messages transmitted only among unbiased agents, the likelihood that it was initially created by a biased agent is much lower, namely $b_{U1} = 0.250$. Therefore, relying on their relational connections to others they know to be trustworthy, users can safely transmit true rumors and rectifications.

The percentage of biased agents within this scenario is highest across all scenarios: on average, only about 8% of the agents are unbiased. This implies, first, that although networked communications exchanges might support unbiased agents in spreading truthful messages amongst themselves, false rumors might still prevail within the population as a whole. Second, if overall there are only few unbiased agents from whom information is available, the relative influence of these agents could be disproportionately high. In fact, the parameter configurations identified could facilitate at least moderately centralized patterns of information flow in which emergent beliefs might be biased towards the beliefs of a few influential agents (Golub and Jackson 2010).

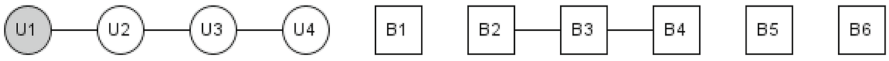

Fig. 4 Rumor transmission along agents' relational ties enables unbiased agents to communicate trustworthy messages amongst themselves

6 Discussion and conclusion

If users learn about a rumor circulating in SMN through dedicated platform features for accessing digital content (e.g., keyword searches and automated information streams such as trending topics), rather than through interacting with others with whom they share a relational connection (e.g., with whom they are friends or whom they follow on a platform), it can divert the flow of social influence among them and hence affect how they evaluate rumor veracity. Inspection of five scenarios of rumor propagation on Twitter reveals that the primary mechanism through which searchability can facilitate self-correction is *motivating users to block unverified messages regardless of their veracity*. If users are generally suspicious of rumors, they may simply choose not support them because of an unwillingness to risk that they might turn out to be false. On the one hand, this can prevent false rumors from spreading; on the other, it makes bottom-up self-correction difficult to achieve. This insight could help explain why statements from official sources and professional journalists play a major role in enabling self-correction (e.g., Andrews et al. 2016; Starbird et al. 2018), and why users may have difficulties distinguishing between true and false rumors a priori, as claimed by Zubiaga et al. (2016a).

In addition, the results point to three mechanisms through which communication among users who share a relational connection can facilitate identifying true rumors. First, such interactions can *create trustworthy transmission pathways for individual messages* transmitted through a sequence of (mostly) trustworthy users. Second, they can *create a communication environment in which truth-seeking agents are willing to spread rumor messages*, as shifts in the relative share of agents who potentially spread false rumors within and outside their connected component can alter their assessment of potential rumor sources' trustworthiness. Finally, communication along their relational ties can also *enable truth-seeking agents to spread presumably true rumor messages amongst themselves*. If trustworthy agents connect primarily with similar others, it could allow them to keep spreading a rumor, regardless of the communication among the rest of the population.

These insights suggest that when users cannot verify the information conveyed by a message itself, it is crucial that they be able to evaluate the trustworthiness of those who have transmitted it. However, empirical research suggests that the interactions between actors who maintain a relational connection play a minor role in the retransmission of rumor messages (e.g., Carlos et al. 2013; Kwon et al. 2017; Vosoughi et al. 2018). It is therefore unlikely that interpersonal trust regarding others' rumor-spreading behaviors and intentions will emerge through social enforcement and network closure, as claimed, for instance, by Coleman (1988) and Granovetter (1985).

However, trust does not result only from agents' social relationships, but also from their preexisting dispositions, social norms, the threat of formal sanctions or reputational damage, and role expectations (Droege et al. 2003). Furthermore, as Burt (2001) points out, trust can facilitate informational benefits that result from network brokerage rather than closure. In the context of SMN, Grabner-Kräuter and Bitter (2015) argue that it is users' *initial trust* in others, based on their foremost

perceptions of trust-relevant attributes (e.g., others' competence, benevolence, honesty, and predictability; McKnight et al. 1998), that might encourage them to acquire information from others. SMN can support the formation of initial trust among users by providing them with cues that help them evaluate the trustworthiness of an information source, such as their connectedness, authority status, identifiability, and others' recommendations of digital content created by them (e.g., Lin et al. 2016; Westerman et al. 2012; Winter et al. 2016). In line with that, Kim et al. (2019) suggest that features that allow users to rate each other might help prevent the spread of content created by untrustworthy actors—even though a prominent finding of their study is that *confirmation bias* can still prevent users from adopting truthful information that is inconsistent with their preexisting beliefs.

One way out of this might be to move toward more institutionalized forms of trust that do not require users to establish relational ties a priori. Turcotte et al. (2015), for instance, suggest that recommendations from perceived opinion leaders can increase other users' trust in digital content. This implies, however, that those opinion leaders would have to acquire a credible reputation a priori. This approach is currently pursued by websites such as *Snopes.com* and journalists who try to establish themselves as fact-checking institutions in SMN as well as in other online and offline channels. Similarly, social bots might serve as trustworthy intermediaries that enable users to distinguish between true and false rumors more easily (e.g., Ciampaglia 2018; Mønsted et al. 2017).

Furthermore, the findings suggest that apart from users' mere connectivity, sufficiently high levels of *homophily* are crucial to ensure that those actors who seek the truth behind a rumor primarily receive trustworthy information from their direct and indirect contacts. While homophily does not ensure that all actors will arrive at adequate veracity judgments, it allows truth-seeking agents to establish trustworthy transmission pathways for particular messages and to communicate such information amongst themselves.

However, a tendency to connect to those who hold similar information preferences can also lead to the emergence of so-called *echo chambers*—polarized communities of actors that hold common views but who are unlikely to adopt information shared by others outside their communities (e.g., Bakshy et al. 2015; Del Vicario et al. 2016; Schmidt et al. 2017). The same mechanisms that can enable agents to focus on trustworthy information hence imply that they might disengage from the general discourse. Thus, even if unbiased agents arrive at adequate veracity judgments, it does not necessarily imply a consensual outcome. On the contrary, false rumors may persist within different communities or may even be held by a majority of users.

Using ABS, I have thus disclosed qualitative tendencies, but not quantified the presumed the relationship between searchability and rumor self-correction in SMN, which is a major limitation of this paper. The propagation scenarios identified are based on input parameter configurations that can reproduce empirically observed patterns of rumor propagation on Twitter, but they do not actually describe users' behaviors, and having identified certain parameter ranges a given number of times does not correspond to a meaningful percentage of observations. For instance, while four of the five identified scenarios point to mechanisms

through which searchability might impede self-correction, knowing that does not allow for predicting its occurrence. What is more, outcomes can be interpreted validly only within the boundaries of the theoretical model, which is an abstract and simplified representation of the real-world social system. While the study thus yields insights into the existence and direction of effects, measurement and falsification necessarily lie within the domain of empirical research rather than that of simulation models.

Furthermore, empirical validation is a major issue when using ABS to identify and illustrate causal mechanisms (Beese et al. 2019). The present study relies on indirect calibration because restrictions in the availability of secondary data have prevented validating the micro-level model constructs directly. The empirical validity of a model that is in qualitative agreement with macro-level patterns of observations is, however, naturally limited (Fagiolo et al. 2019). Furthermore, the insights are contingent on the operating logics of Twitter and the diffusion of rumors pertaining to breaking news events. To some extent, further empirical studies—for instance, case studies or field experiments that collect data from different platforms or that employ longitudinal analyses of communication patterns—could increase confidence in the theoretical mechanisms at work. Beyond that, additional validation steps could be taken to curtail input parameter ranges even further. Still, typification models such as used in this paper can help increase our understanding of the emergent properties of social processes (Boero and Squazzoni 2005). Furthermore, this paper focuses on a theoretical explanation rather than an empirical description or prediction, and so it makes sense to confine to structural validation and empirical calibration (Manson 2003).

In the ongoing debate surrounding collective intelligence and the wisdom of crowds in SMN, this work provides insights into how the affordances of popular and widely used platforms might affect the flow of social influence and, consequently, affect collective sense-making. Network dynamics are certainly only one of several factors that promote the propagation of false rumors in SMN, along with, for instance, individual actors' cognitive limitations (e.g., Mannes 2009) and normative social influences (e.g., Nolan et al. 2008); still, understanding the social dynamics that can emerge from people's interdependent interactions is crucial for explaining social outcomes (Lorenz and Neumann 2018).

The collective sense-making processes that can lead to the manifestation of self-correction in SMN have previously been interpreted in terms of a Habermasian public sphere (Fuchs 2017), yet whether SMN sufficiently enable rational deliberation among their users is the subject of an ongoing dispute (Dahlberg 2011). The normative evaluation of the implications of ICT affordances on online deliberation and public opinion is not the focus of this paper, however, and must hence be left to future research. By way of an epistemic justification, Gelfert (2013) explains that rumors may convey information to actors that they could not otherwise obtain from official sources, which is in line with the more general claim that they can provide people with what might be relevant information about an uncertain and potentially threatening event or situation (DiFonzo and Bordia 2007a). As Coady (2012, p 92) puts it, "no rumor could survive *as a rumor* […] if most of those spreading it were completely indifferent to whether it was true." However, users spreading rumors for

reasons other than their presumed veracity can be epistemologically problematic (Goldman 1995). The present study has referred to actors who might be willing to spread false rumors for desires or preferences that are other than truth-related as *biased* agents. While their motives to spread rumors might be as honorable as those of unbiased agents, they still may take advantage of truth-seeking agents' uncertainty to promote the spread of misinformation in SMN.

Funding Open Access funding enabled and organized by Projekt DEAL.

Code availability The NetLogo code is available from the author by request.

References

Abdullah NA, Nishioka D, Tanaka Y, Murayama Y (2017) Why I retweet? Exploring user's perspective on decision-making of information spreading during disasters. In: Proceedings of the 50th Hawaii International Conference on system sciences, pp 432–441

Allcott H, Gentzkow M (2017) Social media and fake news in the 2016 election. J Econ Perspect 31:211–236

Andrews C, Fichet E, Ding Y, Spiro ES, Starbird K (2016) Keeping up with the tweet-dashians: the impact of "Official" accounts on online rumoring. In: Proceedings of the 19th ACM Conference on computer-supported cooperative work and social computing, pp 452–465

Arif A, Robinson JR, Stanek SA, Fichet E, Townsend P, Worku Z, Starbird K (2017) A closer look at the self-correcting crowd: examining corrections in online rumors. In: Proceedings of the 20th ACM Conference on computer supported cooperative work and social computing, pp 155–168

Bakshy E, Messing S, Adamic LA (2015) Exposure to ideologically diverse news and opinion on Facebook. Science 348:1130–1132

Baym NK, Boyd D (2012) Socially mediated publicness: an introduction. J Broadcast Electron Media 56:320–329

Beese J, Haki MK, Aier S, Winter R (2019) Simulation-based research in information systems. Bus Inf Syst Eng 61:503–521

Berger K, Klier J, Klier M, Probst F (2014) A review of information systems research on online social networks. Commun Assoc Inf Syst 35:8

Bloch F, Demange G, Kranton R (2018) Rumors and social networks. Int Econ Rev 59:421–448

Boero R, Squazzoni F (2005) Does empirical embeddedness matter? Methodological issues of agent-based models for analytical social science. J Artif Soc Soc Simul 8:6

Bordia P, DiFonzo N (2004) Problem solving in social interactions on the internet: rumor as social cognition. Soc Psychol Q 67:33–49

Boyd D (2011) Social network sites as networked publics: affordances, dynamics, and implications. In: Papacharissi Z (ed) A networked self: identity, community, and culture on social network sites. Routledge, New York, pp 39–58

Boyd D, Golder S, Lotan G (2010) Tweet, tweet, retweet: conversational aspects of retweeting on Twitter. In: Proceedings of the 43rd Hawaii International Conference on system sciences, pp 1–10
Burt RS (2001) Bandwidth and echo: trust, information, and gossip in social networks. In: Rauch JE, Casella A (eds) Networks and markets. Russell Sage Foundation, New York, pp 30–74
Buskens V (1998) The social structure of trust. Soc Netw 20:265–289
Carlos C, Mendoza M, Poblete B (2013) Predicting information credibility in time-sensitive social media. Internet Res 23:560–588
Cialdini RB, Goldstein NJ (2004) Social influence: compliance and conformity. Annu Rev Psychol 55:591–621
Ciampaglia GL (2018) Fighting fake news: a role for computational social science in the fight against digital misinformation. J Comput Soc Sci 1:147–153
Coady D (2012) What to believe now: applying epistemology to contemporary issues. Wiley-Blackwell, Malden
Coleman JS (1988) Social capital in the creation of human capital. Am J Sociol 94:S95–S120
Dahlberg L (2011) Re-constructing digital democracy: an outline of four "Positions." New Media Soc 13:855–872
Davis JP, Eisenhardt KM, Bingham CB (2007) Developing theory through simulation methods. Acad Manag Rev 32:480–499
Del Vicario M, Vivaldo G, Bessi A, Zollo F, Scala A, Caldarelli G, Quattrociocchi W (2016) Echo chambers: emotional contagion and group polarization on Facebook. Sci Rep 6:37825
DeMarzo PM, Vayanos D, Zwiebel J (2003) Persuasion bias, social influence, and unidimensional opinions. Q J Econ 118:909–968
Denrell J, Le Mens G (2017) Information sampling, belief synchronization, and collective illusions. Manag Sci 63:528–547
DiFonzo N, Bordia P (2007a) Rumor, gossip and urban legends. Diogenes 54:19–35
DiFonzo N, Bordia P (2007b) Rumor psychology: social and organizational approaches. American Psychological Association, Washington, DC
Doerr B, Fouz M, Friedrich T (2012) Why rumors spread so quickly in social networks. Commun ACM 55:70–75
Droege SB, Anderson JR, Bowler M (2003) Trust and organizational information flow. J Bus Manag 9:45–59
Ellison NB, Gibbs JL, Weber MW (2015) The use of enterprise social network sites for knowledge sharing in distributed organizations: the role of organizational affordances. Am Behav Sci 59:103–123
Elsweiler D, Harvey M (2015) Engaging and maintaining a sense of being informed: understanding the tasks motivating Twitter search. J Assoc Inf Sci Technol 66:264–281
Evans SK, Pearce KE, Vitak J, Treem JW (2017) Explicating affordances: a conceptual framework for understanding affordances in communication research. J Comput-Mediat Commun 22:35–52
Fagiolo G, Guerini M, Lamperti F, Moneta A, Roventini A (2019) Validation of Agent-Based Models in Economics and Finance. In: Beisbart C, Saam NJ (eds) Computer Simulation Validation: Fundamental Concepts, Methodological Frameworks, and Philosophical Perspectives. Springer, Cham, pp 763–787
Friggeri A, Adamic LA, Eckles D, Cheng J (2014) Rumor cascades. In: Proceedings of the 8th International AAAI Conference on weblogs and social media, pp 101–110
Fuchs C (2017) Social media: a critical introduction, 2nd edn. Sage, London
Garg R, Smith MD, Telang R (2011) Measuring information diffusion in an online community. J Manag Inf Syst 28:11–37
Garton L, Haythornthwaite C, Wellman B (1997) Studying online social networks. J Comput Mediat Commun 3:JCMC313
Gelfert A (2013) Coverage-reliability, epistemic dependence, and the problem of rumor-based belief. Philosophia 41:763–786
Goldman AI (1995) Social epistemology, interests, and truth. Philos Top 23:171–187
Golub B, Jackson MO (2010) Naïve learning in social networks and the wisdom of crowds. Am Econ J Microecon 2:112–149
Grabner-Kräuter S, Bitter S (2015) Trust in online social networks: a multifaceted perspective. Forum Soc Econ 44:48–68
Granovetter M (1985) Economic action and social structure: the problem of embeddedness. Am J Sociol 91:481–510

Gupta A, Lamba H, Kumaraguru P, Joshi A (2013) Faking sandy: characterizing and identifying fake images on Twitter during Hurricane Sandy. In: Proceedings of the 22nd International Conference on World Wide Web, pp 729–736

Hannak A, Margolin D, Keegan B, Weber I (2014) Get back! You don't know me like that: the social mediation of fact checking interventions in twitter conversations. In: Proceedings of the 8th International AAAI Conference on weblogs and social media, pp 187–196

Hart PS, Nisbet EC (2012) Boomerang effects in science communication: how motivated reasoning and identity cues amplify opinion polarization about climate migration policies. Commun Res 39:701–723

Heverin T, Zach L (2012) Use of microblogging for collective sense-making during violent crises: a study of three campus shootings. J Am Soc Inf Sci Technol 63:34–47

Honeycutt C, Herring SC (2009) Beyond microblogging: conversation and collaboration via Twitter. In: Proceedings of the 42nd Hawaii International Conference on system sciences

Huang J, Thornton KM, Efthimiadis EN (2010) Conversational tagging in Twitter. In: Proceedings of the 21st ACM Conference on hypertext and hypermedia, pp 173–178

Jong W, Dückers MLA (2016) Self-correcting mechanisms and echo-effects in social media: an analysis of the "Gunman in the Newsroom" Crisis. Comput Hum Behav 59:334–341

Kane GC, Alavi M, Labianca G, Borgatti SP (2014) What's different about social media networks?: a framework and research agenda. Manag Inf Syst Q 38:275–304

Karahanna E, Xu SX, Xu Y, Zhang N (2018) The needs-affordances-features perspective for the use of social media. Manag Inf Syst Q 42:737–756

Kim A, Moravec PL, Dennis AR (2019) Combating fake news on social media with source ratings: the effects of user and expert reputation ratings. J Manag Inf Syst 36:931–968

Kimmel AJ, Audrain-Pontevia A (2010) Analysis of commercial rumors from the perspective of marketing managers: rumor prevalence, effects, and control tactics. J Mark Commun 16:239–253

Kleijnen JPC (1995) Verification and validation of simulation models. Eur J Oper Res 82:145–162

Klein D, Marx J, Fischbach K (2018) Agent-based modeling in social science, history, and philosophy: an introduction. Hist Soc Res 43:7–27

Kogan M, Palen L (2018) Conversations in the eye of the storm: at-scale features of conversational structure in a high-tempo, high-stakes microblogging environment. In: Proceedings of the 2018 CHI Conference on human factors in computing systems, paper 84

Kwon KH, Bang CC, Egnoto M, Rao HR (2016) Social media rumors as improvised public opinion: semantic network analyses of twitter discourses during Korean Saber Rattling 2013. Asian J Commun 26:201–222

Kwon S, Cha M, Jung K (2017) Rumor detection over varying time windows. PLoS ONE 12:e0168344

Lee J, Agrawal M, Rao HR (2015) Message diffusion through social network service: the case of rumor and non-rumor related tweets during boston bombing 2013. Inf Syst Front 17:997–1005

Leidner DE, Gonzalez E, Koch H (2018) An affordance perspective of enterprise social media and organizational socialization. J Strateg Inf Syst 27:117–138

Lerman K, Ghosh R (2010) Information contagion: an empirical study of the spread of news on Digg and Twitter social networks. In: Proceedings of the 4th International AAAI Conference on weblogs and social media, pp 90–97

Leskovec J, Lang KJ, Dasgupta A, Mahoney MW (2008) Statistical properties of community structure in large social and information networks. In: Proceedings of the 17th International Conference on World Wide Web, pp 695–704

Lewandowsky S, Ecker UKH, Seifert CM, Schwarz N, Cook J (2012) Misinformation and its correction: continued influence and successful debiasing. Psychol Sci Public Interest 13:106–131

Li B, Chong AY (2019) What influences the dissemination of online rumor messages: message features and topic-congruence. In: Proceedings of the 40th International Conference on information systems

Li H, Sakamoto Y (2014) Social impacts in social media: an examination of perceived truthfulness and sharing of information. Comput Hum Behav 41:278–287

Li H, Sakamoto Y, Tanaka Y, Chen R (2014) The Psychology behind People's Decision to Forward Disaster-Related Tweets. In: Proceedings of the 19th Pacific Asia Conference on information systems, p 123

Lin X, Spence PR, Lachlan KA (2016) Social media and credibility indicators: the effect of influence cues. Comput Hum Behav 63:264–271

Lorenz J, Neumann M (2018) Opinion dynamics and collective decisions. Adv Complex Syst 21:1802002
Macy MW, Willer R (2002) From factors to actors: computational sociology and agent-based modeling. Annu Rev Sociol 28:143–166
Maddok J, Starbird K, Al-Hassani H, Sandoval DE, Orand M, Mason RM (2015) Characterizing online rumoring behavior using multi-dimensional signatures. In: Proceedings of the 18th ACM Conference on computer supported cooperative work and social computing, pp 228–241
Mannes AE (2009) Are we wise about the wisdom of crowds? The use of group judgments in belief revision. Manag Sci 55:1267–1279
Manson SM (2003) Validation and verification of multi-agent models for ecosystem management. In: Janssen M (ed) Complexity and ecosystem management: the theory and practice of multi-agent approaches. Edward Elgars Publishers, Northampton, pp 63–74
Margolin DB, Hannak A, Weber I (2018) Political fact-checking on twitter: when do corrections have an effect? Polit Commun 35:196–219
Marino S, Hogue IB, Ray CJ, Kirschner DE (2008) A methodology for performing global uncertainty and sensitivity analysis in systems biology. J Theor Biol 245:178–196
Mason WA, Conrey FR, Smith ER (2007) Situating social influence processes: dynamic, multidirectional flows of influence within social networks. Pers Soc Psychol Rev 11:279–300
McKnight DH, Cummings LL, Chervany NL (1998) initial trust formation in new organizational relationships. Acad Manag Rev 23:473–490
McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: homophily in social networks. Annu Rev Sociol 27:415–444
Melamed D, Breiger RL, Schoon E (2012) The Duality of Clusters and Statistical Interactions. Sociol Methods Res 42:41–59
Mønsted B, Sapieżyński P, Ferrara E, Lehmann S (2017) Evidence of complex contagion of information in social media: an experiment using Twitter Bots. PLoS One 12:e0184148
Nolan JM, Wesley Schultz P, Cialdini RB, Goldstein NJ, Griskevicius V (2008) Normative social influence is underdetected. Pers Soc Psychol Bull 34:913–923
Nyhan B, Reifler J (2010) When corrections fail: the persistence of political misperceptions. Polit Behav 32:303–330
Oh O, Agrawal M, Rao HR (2013) Community intelligence and social media services: a rumor theoretic analysis of tweets during social crises. Manag Inf Syst Q 37:407–426
Oh O, Eom C, Rao HR (2015) Role of social media in social change: an analysis of collective sense making during the 2011 Egypt Revolution. Inf Syst Res 26:210–223
Pfeffer J, Zorbach T, Carley KM (2014) Understanding online firestorms: negative word-of-mouth dynamics in social media networks. J Mark Commun 20:117–128
Procter R, Vis F, Voss A (2013) Reading the riots on Twitter: methodological innovation for the analysis of big data. Int J Soc Res Methodol 16:197–214
R Core Team (2019) R: a language and environment for statistical computing, R Foundation for Statistical Computing, Vienna. https://www.R-project.org/. Accessed 4 Sept 2020
Salecker J, Sciaini M, Meyer KM, Wiegand K (2019) The nlrx R package: a next-generation framework for reproducible NetLogo. Model Anal 10:1854–1863
Schmidt AL, Zollo F, Del Vicario M, Bessi A, Scala A, Caldarelli G, Stanley EH, Quattrociocchi W (2017) Anatomy of news consumption on Facebook. Proc Natl Acad Sci USA 114:3035–3039
Serrano E, Iglesias CE, Garijo M (2015) A survey of twitter rumor spreading simulations. In: Núñez M, Nguyen NT, Camacho T, Trawiński B (eds) Computational collective intelligence: 7th International Conference, Madrid, Spain, September 21–23, 2015, Proceedings. Part I, pp 113–122
Shin J, Jian L, Driscoll K, Bar F (2017) Political rumoring on Twitter during the 2012 US Presidential Election: rumor diffusion and correction. New Media Soc 19:1214–1235
Shin J, Jian L, Driscoll K, Bar F (2018) the diffusion of misinformation on social media: temporal pattern, message, and source. Comput Hum Behav 83:278–287
Sîrbu A, Loreto V, Servedio VDP, Tria F (2017) Opinion dynamics: models, extensions and external effects. In: Loreto V, Haklay M, Servedio VDP, Stumme G, Theunis J, Tria F (eds) Participatory sensing, opinions and collective awareness. Springer, Cham, pp 363–401
Smith ER, Conrey FR (2007) Agent-based modeling: a new approach for theory building in social psychology. Pers Soc Psychol Rev 11:87–104

Spiro ES, Sutton J, Greczek M, Fitzhugh S, Pierski N, Butts CT (2012) Rumoring during extreme events: a case study of deepwater horizon 2010. In: Proceedings of the 4th Annual ACM Web Science Conference, pp 275–283

Starbird K, Maddok J, Orand M, Achterman P, Mason RM (2014) Rumors, false flags, and digital vigilantes: misinformation on Twitter after the 2013 Boston Marathon Bombing. In: iConference Notes, pp 654–662

Starbird K, Dailey D, Mohamed O, Lee G, Spiro ES (2018) Engage early, correct more: how journalists participate in false rumors online during crisis events. In: Proceedings of the CHI Conference on human factors in computing systems

Stieglitz S, Bunker D, Mirbabaie M, Ehnis C (2018) Sense-making in social media during extreme events. J Conting Crisis Manag 26:4–15

Stonedahl F, Wilensky U (2010) BehaviorSearch, Center for Connected Learning and Computer Based Modeling, Northwestern University, Evanston, IL. http://www.behaviorsearch.org. Accessed 4 Sept 2020

Takayasu M, Sato K, Sano Y, Yamada K, Miura W, Takayasu H (2015) Rumor diffusion and convergence during the 3.11 earthquake: a Twitter Case Study. PLoS ONE 10:e0121443

Teevan J, Ramage D, Morris MR (2011) #TwitterSearch: a comparison of microblog search and web search. In: Proceedings of the 4th ACM International Conference on web search and data mining, pp 35–44

Thiele JC, Kurth W, Grimm V (2014) Facilitating parameter estimation and sensitivity analysis of agent-based models: a cookbook using NetLogo and R. J Artif Soc Soc Simul 17:11

Turcotte J, York C, Irving J, Scholl RM, Pingree RJ (2015) News recommendations from social media opinion leaders: effects on media trust and information seeking. J Comput-Mediat Commun 20:520–535

Valkenburg PM, Peter J, Walther JB (2016) Media effects: theory and research. Annu Rev Psychol 67:315–338

Vosoughi S, Roy D, Aral S (2018) The spread of true and false news online. Sci 359:1146–1151

Wang B, Zhuang J (2018) Rumor response, debunking response, and decision makings of misinformed twitter users during disasters. Nat Hazards 93:1145–1162

Wasserman S, Faust K (1994) Social network analysis: methods and applications. Cambridge University Press, Cambridge

Westerman P, Spence PR, van der Heide B (2012) A social network as information: the effect of system generated reports of connectedness on credibility on Twitter. Comput Hum Behav 28:199–206

Wilensky U (1999) NetLogo, Center for Connected Learning and Computer-Based Modeling, Northwestern University. Evanston, IL. http://ccl.northwestern.edu/netlogo. Accessed 4 Sept 2020

Windrum P, Fagiolo G, Moneta A (2007) Empirical validation of agent-based models: alternatives and prospects. J Artif Soc Soc Simul 10:8

Winter S, Metzger MJ, Flanagin AJ (2016) Selective use of news cues: a multiple-motive perspective on information selection in social media environments. J Commun 66:669–693

Young HP (2009) Innovation diffusion in heterogeneous populations: contagion, social influence, and social learning. Am Econ Rev 99:1899–1924

Zannettou S, Sirivianos M, Blackburn J, Kourtellis N (2019) The web of false information: rumors, fake news, hoaxes, clickbait, and various other Shenanigans. J Data Inf Qual 11:10

Zhang B, Pavlou PA, Krishnan R (2018) On direct vs. Indirect peer influence in large social networks. Inf Syst Res 29:292–314

Zhao L, Yin J, Song Y (2016) An exploration of rumor combating behavior on social media in the context of social crises. Comput Hum Behav 58:25–36

Zollo F, Bessi A, Del Vicario M, Scala A, Caldarelli G, Shekhtman L, Havlin S, Quattrociocchi W (2017) Debunking in a world of Tribes. PLoS ONE 12:e0181821

Zubiaga A, Liakata M, Procter R, Bontcheva K, Tolmie P (2015) Crowdsourcing the annotation of rumourous conversations in social media. In: Proceedings of the 24th International Conference on World Wide Web, pp 347–353

Zubiaga A, Liakata M, Procter R, Wong Sak Hoi G, Tolmie P (2016a) Analysing how people orient to and spread rumours in social media by looking at conversational threads. PLoS ONE 11:e0150989

Zubiaga A, Liakata M, Procter R, Wong Sak Hoi G, Tolmie P (2016b) PHEME rumour scheme dataset: journalism use case. https://figshare.com/articles/PHEME_rumour_scheme_dataset_journalism_use_case/2068650. Accessed 4 Sept 2020

Zubiaga A, Aker A, Bontcheva K, Liakata M, Procter R (2018) Detection and resolution of rumours in social media: a survey. ACM Comp Surv 51:32