If you see something, then say something to others: Complex contagions and the

socially-contingent correction of misinformation

Hyunjin Song

Department of Communication, University of Vienna, Austria

Word count: 6630 words

Draft date: March 29, 2018

Draft in progress. Please do not cite without permission.

Please direct any questions and inquiries to hyunjin.song@univie.ac.at

Author Note

Abstract

As the citizens' news consumption is increasingly driven by online sources, the propagation of misinformation and so-called "fake news" on those platforms become an increasing concern for the public and policy makers. Our goal in this contribution is to offer a more systematic assessment of underlying mechanisms of misinformation spreading and its correction, combining a macro social contextual factor and individuals' cognitive basis of adopting misinformation into a more integrated, dynamic system model perspective. We first review existing evidence concerning individuals' cognitive basis of adopting such misinformation, and social context of which exposure to misinformation and its corrections are received. Next, adopting a well-known class of an epidemic model of virus infection and recovery, we combine this micro and macro dynamics into comprehensive, integrated model of misinformation diffusion on social networks. We do so by focusing on the distinction between simple contagion of misinformation vs. complex contagion of adopting corrective messages. Relying on Agent-based simulations, we further explore various boundary conditions of such dynamics, aiming to uncover how and when such misinformation propagates into the public, as well as what factors facilitate or hinder such diffusion process.

Keywords: Misinformation, fake news, correction, simple contagion, complex contagion, exponential random graph model, agent-based simulations

If you see something, then say something to others: Complex contagions and the

socially-contingent correction of misinformation

Citizens across the worlds are experiencing major changes in their news environments with the development of digital media. One of the most dramatic changes in the news environment in recent decades involves the role social networking sites (SNS) such as Facebook and Twitter play as a primary source of news outlets. Not only citizens' news consumptions are increasingly driven by such online sources (Shearer & Gottfried, 2017), but it also appears that citizens themselves are actively participating in news dissemination on those platforms by sharing news with their peers (e.g., Lee & Song, 2017; Shearer & Gottfried, 2017).

An effective deliberation among public is regarded as a keystone of thriving democracies, and modern political systems squarely depend on informed decisions of citizens in that regard (Delli Carpini & Keeter, 1996). Yet, a propagation of rumors, misinformation, and so-called "fake news" on those platforms becomes an increasing concern for the public and policy makers alike (Allcott & Gentzkow, 2017; Lazer et al., 2017), as evidenced in recent 2016 U.S. presidential election (Allcott & Gentzkow, 2017; Giglietto, Iannelli, Rossi, & Valeriani, 2016; Guess, Nyhan, & Reifler, 2018) and in Brexit votes (The New York Times, 2017). While a wide circulation of factually dubious information is not entirely new to political arena, a growing trend of digitally disseminated rumors and misinformations – often termed as a "fake news" phenomenon – is increasingly recognized as a serious threat to liberal democratic societies (Allcott & Gentzkow, 2017; Lazer et al., 2017). Either based on unsubstantiated rumors or based on factually wrong beliefs, many of the misinformed behave differently than those who are accurately informed (Kuklinski, Quirk, Jerit, Schwieder, & Rich, 2000). They often disagree about basic facts about many public issues, and continue to believe and rely on such false information when making political judgments (Nyhan & Reifler, 2010; Thorson, 2016).

Along with these trends, there has been an growing interest among scholars on how people process and maintain factually false (or at least factually dubious) information from the perspectives of an individual's cognitive processes (Kuklinski et al., 2000; Lewandowsky, Ecker, Seifert, Schwarz, & Cook, 2012; Weeks, 2015). These studies have generated a valuable insights of how individuals often maintain factually false beliefs, and further, how corrections to

false beliefs are received and processed (Garrett, Weeks, & Neo, 2016; Lewandowsky et al., 2012; Thorson, 2016). However, despite growing interest and continued research effort to better understand the nature and its exact mechanism, what we know about the spread of misinformation and fake news specifically on online social networks is largely based on limited evidence due to its complex nature of the problem.

Against this backdrop, our goal in this contribution is to offer a more systematic assessment of underlying mechanisms of misinformation spreading and its correction, focusing on one's *social contexts* in which such (mis)information and corrective messages are received and processed. We argue that while *exposure* to (mis)information is likely to follow a simple contagion process, *changes* in one's beliefs regarding such (mis)information – which ultimately *the* goal of corrective messages – is likely to be, in Centola and Macy's (2007) term, a "complex contagion" where such changes require multiple sources of affirmation and reinforcement compared to simple contagion process. As a result, the effects of fact-checking and corrective messages are likely to be highly *socially* contingent, yet studies only now begin to consider this possibility more seriously (Bode & Vraga, 2017; Margolin, Hannak, & Weber, 2017).

In what follows, we first review existing evidence regarding political misperceptions and the effects of fact-checking (i.e., correction) messages. We advance our perspective by combining an individual-level cognitive and affective basis of adopting such misinformation with a social context of which misinformation and corrections are received. Based on a well-known class of an epidemic model of virus infection and recovery, we propose an integrated model of misinformation diffusion and socially-contingent corrections on social networks, with a special focus on the differences between a *simple contagion* of misinformation and a *complex contagion* of corrections and fact-checking messages. Relying on Agent-Based Model (ABM) simulations, we robustly explore boundary conditions of such dynamics, aiming to uncover how and when such misinformation propagates into the public.

Psychological Underpinnings of Fake News, Misperceptions, and Corrections

Following Allcott and Gentzkow's (2017) definition, we define *fake news* as "distorted signals uncorrelated with the truth" (p. 212). This encompasses several related concepts, such as misinformation, rumors, and disinformation. Literature on this topic generally maintain loosely

defined, but at the same time highly interrelated, conceptualizations of those related terms. For instance, (political) rumors are often defined as "unsubstantiated claims about candidates and issues that are often false" (Weeks & Garrett, 2014, p. 401). Similarly, misinformation (or misperceptions) are defined as factual information (or beliefs) "that are false or contradict the best available evidence in the public domain" (Flynn, Nyhan, & Reifler, 2017, p. 128). In relation to this, *dis*information campaigns often denote organized, strategic efforts that trying to sway public opinion using rumors and misinformation (Garrett, 2017; Lewandowsky et al., 2012). Understood in this way, fake news often exclude unintentional reporting mistakes, parodies and satires, or unverifiable conspiracy theories (Allcott & Gentzkow, 2017). While term *fake news* often than not additionally entail specific pseudo-journalistic styles that mimic legitimate news sources to intentionally deceiving audiences (Egelhofer & Lecheler, 2017), we use term "fake news" somewhat loosely, denoting any type of misinformation – information that is not supported by best-available evidence – that is deliberatively circulated among publics.[1]

Literature on misinformation and its persistence often converges to the observation that publics' exposure to and acceptance of misinformation are largely driven by one's motivated consistency needs. That is, people disproportionately gravitate toward information that conforms to their partisan priors (Garrett et al., 2016; Weeks & Garrett, 2014), and more likely to accept and endorse such messages (Guess et al., 2018; Nyhan & Reifler, 2010). A mounting evidence – largely based on Kunda's (1990) or on Taber and Lodge's (2006) motivated reasoning framework – suggests that citizens tend to evaluate attitudinally congruent information as more convincing and valid *regardless of its veracity*, while attitudinally inconsistent information is likely to be perceived as weak and therefore likely to be rejected. Therefore, it is perhaps not surprising to find that most of the prior studies based on a motivated reasoning framework document that fact-checking messages (sometimes denoted as "corrective" or "debunking" messages in the literature) have only limited effects due to inherent tendency of humans to directionally process politically relevant information (Flynn et al., 2017; Taber & Lodge, 2006; Thorson, 2016). Even worse, corrective messages may backfire, may induce higher level of endorsements of false beliefs than actually lower them (e.g., Nyhan & Reifler, 2010; but see Wood & Porter, 2018).

Another line of studies based on a dual process theory of human cognitive processing and memory suggests that attitudinally-congruent misinformation tends to form a coherent mental model based on one's partisan schema and stereotypes (e.g., Garrett, Nisbet, & Lynch, 2013; Lewandowsky et al., 2012). Therefore, people tend to fill any gaps caused by corrections (that invalidate some parts of the existing mental model) with flawed but attitudinally congruent misinformation that is still readily accessible in their memory (Lewandowsky et al., 2012). Studies also find that this effect is much more likely when correction messages do not update the initial mental model that justifies misinformation (Chan, Jones, Jamieson, & Albarracín, 2017), when the perceived veracity of initial misinformation is high (due to fluency bias in one's cognitive processing: Lewandowsky et al., 2012), or when individuals can generate counter-arguing reasons in support for initial misperceptions (Chan et al., 2017; Garrett et al., 2013). Most importantly, due to limitations of strategic memory search processes (i.e., it requires more effortful processing), people may still rely on negated misinformation in subsequent reasoning *even when they remember such information is factually incorrect* (Lewandowsky et al., 2012). Therefore, even in the face of seemingly effective corrections, the effect of misperceptions lingers and continue to exert influence (Thorson, 2016).

Under certain situations, it seems that citizens *indeed* can adhere factual information based on correction messages despite their perpetual partisan bias (e.g., Nyhan, Porter, Reifler, & Wood, 2017; Wood & Porter, 2018). Yet as Margolin et al. (2017) note, it appears that such effects often require special *social context*. This observation is indeed much warranted, as most of the previous studies concerning misinformation and the effect of fact-checking messages are conducted in an experimental context with a single-shot, *asocial* correction message from media professionals and fact-checking organizations (e.g., Garrett et al., 2013; Nyhan & Reifler, 2010; Weeks, 2015). There is also a suggestive evidence that fact-checking and corrective messages from one's peers in their social networks – what we would call a "social correction" – are more likely to be effective in reducing misperceptions (e.g., Bode & Vraga, 2017; Margolin et al., 2017). In what follow, we review several theoretical accounts of such *socially-based* correction messages on misinformation and fake news.

A Social Context of Misinformation and Corrections: Simple vs. Complex Contagions

People's perceptions and behaviors are likely to be shaped by their social contacts (Centola & Macy, 2007; Lazer, Rubineau, Chetkovich, Katz, & Neblo, 2010), and therefore perceptions and behaviors may spread within social networks (e.g., Bond et al., 2012). Indeed, a non-negligible number of prior accounts of partisan misinformation and fake news on social networks connects this idea to possible mechanisms of misinformation propagation and *spreading* (e.g., Bakshy, Rosenn, Marlow, & Adamic, 2012; Del Vicario et al., 2016). The most simplest form of those accounts posits that online social networks provide one of the ideal settings for partisan misinformation and fake news to be spread within such networks. Many of the partisan rumors and fake news tend to be richer in their "novelty" (Wu & Huberman, 2007) regardless of its veracity and informational value. As Giglietto et al.'s (2016) observation suggests, such non-redundant and "novel" information (false and unverified information being one of them) tend to be engaging, and spread faster via weaker ties (Granovetter, 1977). You can easily share and post such news with little to no effort, and your friends on your social networks are easily exposed to such information, in turn they also share such news to their own peers, and so on – creating what is called a *information cascade* of misinformation (Del Vicario et al., 2016). Also, individuals maintain lots of weak ties in social network platforms, which is thought to diversify the information flow. This creates a particularly congenial scenario for a (mis)information propagation and dissemination (Bakshy et al., 2012; Granovetter, 1977). Indeed, Guess et al.'s (2018) investigation of fake news consumption during the 2016 U.S. presidential election suggests that Facebook was likely to be the focal gateway for visiting websites that propagates fake news, while Allcott and Gentzkow's (2017) study also reveals that such dubious stories are indeed widely shared on Facebook during the election.

Often, this process of spreading (mis)information within a social network can be described as a *simple contagion* process – the process of which a single contact with an "infected" individual (in this case, those who spread misinformation to their peers) is sufficient for such misinformation to be spread to another individual (Centola, 2010; Mønsted, Sapieżyński, Ferrara, & Lehmann, 2017; Siegel, 2009). While a spread of any human behavior requires a minimum threshold of one, for the diffusion of information itself – much like

communicable diseases – "the threshold is almost always exactly one" (Centola & Macy, 2007, p. 706; also see Centola, 2010). Once your peer shares news, you become (almost automatically) aware of such news, and you do not require someone else to keep asking about the same news in order to be *aware* of it.

Moreover, due to inherent partisan motivated directionality, the threshold of actually *adapting* such a false yet pro-attitudinal claim (i.e., believing attitudinally congruent misinformation) may also exhibit similar low-threshold properties, although the actual threshold for believing misinformation would be bit higher compared to that of mere exposure and awareness (e.g., Mønsted et al., 2017), and also subject to some individual differences (e.g., Flynn et al., 2017; Weeks, 2015). As previously suggested, people tend to directionally process political information (Taber & Lodge, 2006), and they tend to easily remember and recall attitudinally-consistent information than uncongenial ones. Therefore, at least for citizens who find given misinformation to be congenial to their partisan priors, *adapting* such false yet pro-attitudinal claims also does not require high number of repeated exposure nor independent, multiple exposure to different sources supporting such claims. Indeed, many of the prior empirical studies support this perspective (e.g., Flynn et al., 2017; Garrett et al., 2016; Kuklinski et al., 2000).

In contrast, there are reasons to believe why the effectiveness of fact-checking and corrective messages would be different from that of a simple contagion of misinformation. An adoption of a new perspective that contradicts with one's priors (such as an adoption of counter-attitudinal fact-checking messages) is likely to be, in Centola and Macy's (2007) term, a "complex contagion" where it requires *exposure to multiple sources*, rather than *multiple number of exposures*, endorsing such message in order to be accepted and further spread into a given network. This is because many political attitudes and subsequent actions (such as politically motivated misperceptions) are likely to be deeply rooted in one's social identities and values, therefore changes in one's political beliefs and attitudes require multiple sources of affirmation and reinforcement from multiple *contacts* compared to simple contagion cases (e.g., González-Bailón, 2017; Larson, Nagler, Ronen, & Tucker, 2016; Siegel, 2009).

Moreover, such complex contagion dynamics surrounding correction messages are also

likely to be dependent upon the attitudinal composition of one's local network. For the case of a adoption of fact-checking and corrective messages, the number of one's neighbors who are *not* activated (e.g., those who still *endorse* false belief based on misinformation) tend to discourages their neighbors to adopt a correction, whereas the number of one's neighbors who are already activated (e.g., those who do not believe misinformation anymore) would increase one's susceptibility to adopt the correction given exposure to such correction (i.e., a "*contested*" contagion: Centola & Macy, 2007; also see Friedkin, 2001). Indeed, many social contagions are often perceived to be lack of credibility and legitimacy until adopted by one's neighbors, therefore relative distribution of one's neighbors (in terms of supporters vs. opponents of the adoption) critically influence one's decision to adopt a controversial innovation (Friedkin, 2001; González-Bailón, 2017; Larson et al., 2016). This further means that locally-defined social dynamics may drive specific adoption behaviors of counter-attitudinal information, and consequently, such process would non-trivially interact with a structure and its attitudinal composition of a given network (e.g., Friedkin, 2001). Indeed, there exits a considerable support for this perspective, such that different network topologies (Centola, 2010; Siegel, 2009) or attitudinal makeup of social networks[2] (González-Bailón, 2017; Larson et al., 2016; Levitan & Visser, 2008) may produce different attitudinal and behavioral consequences for social influence. This further implies that some network structures are more prone to generating cascades and adoptions than others.

There are also at least several other reasons why socially-based correction messages, especially from one's peers, might be more effective than a single-shot, *asocial* correction message from more distanced sources (such as fact-checking organizations or mere strangers). First, people may evaluate information coming from their peers to be more credible and trustworthy (e.g., Metzger, Flanagin, & Medders, 2010), and more willing to deliberate with their close social contacts (Morey, Eveland, & Hutchens, 2012). Research suggests that while individuals rather maintain much flexible attitudes as long as their social affiliation goals are met (Levitan, 2017), reputational risks and social accountability of rejecting corrective messages run high for more close social contacts compared to more distant sources such as strangers (Margolin et al., 2017).

Second, previous studies concerning citizens' political discussion network suggest that an individual's network construction is not likely driven by overt partisan considerations (Lazer et al., 2010; Song, 2015), therefore there exists a considerable degree of exposure to disagreement in citizens' everyday political interactions (e.g., Bakshy, Messing, & Adamic, 2015; Morey et al., 2012). Under such a situation, attitudinally heterogeneous networks trigger more systematic processing of available information, which make individuals to be more responsive to argument strengths (Levitan & Visser, 2008) or social utility (Messing & Westwood, 2014), therefore makes them less resistant to corrective messages from their peers.

All in all, prior empirical evidence and theoretical perspectives convincingly suggest that a *simple* contagion of misinformation and a *complex* contagion of corrective messages would exhibit different properties for population-level propagation dynamics. Structurally weak, but bridging ties such as distant contacts in social networks may provide sufficient means for a simple (mis)information – much like communicable disease – to be spread, while ideologically-driven directionally motivated reasoning may provide sufficient psychological grounds for partisans to easily adopt and believe such misinformation. In contrast, an adoption of corrective message – similar to controversial innovations – requires independent and multiple reinforcements from many social contacts due to its counter-attitudinal and "contested" nature, critically dependent upon a structure of network and its attitudinal composition (Centola, 2010; Centola & Macy, 2007; González-Bailón, 2017). Aforementioned perspectives therefore undoubtedly point to the possibility that adoptions of fact-checking messages are likely to be highly socially-contingent, and under certain cases, a social correction would be much more effective than an isolated correction message as typically have been considered in previous experiment contexts (e.g., Garrett et al., 2013; Nyhan & Reifler, 2010).

Observational Challenges in Studying Misperception Within Social Networks

If the diffusion of (mis)information and adoption of corrective messages may non-trivially dependent upon a structure of a given network and its attitudinal composition, then how a typical (online) social network is structured in terms of its topological features, and how pervasive is "attitudinal" homophily, as *the* critical factor determining attitudinal compositions of one's social network? How such structural features affect the overall diffusion dynamics

empirically? This is indeed important questions to ask, yet it is surprisingly difficult to establish convincing evidence of the impact of such factors in complex contagion dynamics using purely observational and experimental approaches.

A frequent and recurring theme for an attitudinal makeup of citizens' social networks and its consequences, especially among general publics, is that most of the citizens today are put into a "echo chamber" or a "filter bubble" that insulate themselves from competing viewpoints and attitude-discrepant information (e.g., Del Vicario et al., 2016; Lewandowsky, Ecker, & Cook, 2017). Yet, as Garrett (2017) puts it, "there is ample evidence that *exposure* [emphasis added] echo chambers are not a typical part of Internet users' experience" (p. 370). Most of citizens are appear to be embedded in sufficiently diverse social networks, showing a substantial level of exposure to political difference in online (e.g., Bakshy et al., 2015; Messing & Westwood, 2014) and in offline (e.g., Huckfeldt, Mendez, & Osborn, 2004). More importantly, those studies do not find compelling evidence that ordinary citizens' network constructions are primarily driven by purposive political homophily that limit their exposure to only congenial political perspectives (Lazer et al., 2010; Song, 2015). This suggests that popular claims of echo chambers or filter bubbles are often overstated.[3]

However, one should also bear in mind that the evidence concerning *exposure* to counter-attitudinal messages does not provide sufficient evidence of how such messages are actually cognitively processed and interpreted under such a situation (e.g., see "engagement echo chamber" discussion in Garrett, 2017; also, see Nyhan et al., 2017). What we know about citizens' exposure to disagreement in their social networks typically has relied on observational evidence, either based on participants' self-reports of their patterns of social interactions concerning their immediate social environment (Huckfeldt et al., 2004) or based on a comprehensive mapping of their interactions in a well-defined, relatively closed social system (e.g., Lazer et al., 2010; Song, 2015). Sometimes, scholars also rely on digitally available communication patterns such as messages posted in online social networks (e.g., Margolin et al., 2017), along with engagement indicators such as "likes" or "shares" (e.g., Bakshy et al., 2015) or based on a detailed digital footprint data such as website access data (e.g., Guess et al., 2018). While such evidence *do* suggests that citizens are indeed frequently exposed to competing

political perspectives in their daily lives, such evidence is typically rather silent about how
individuals actually process and interpret such information given exposure. As experimental
evidence to date suggests, the mere presence of ideologically diverse *exposure* does not
automatically translate into the possibility of more balanced judgments. Experimental evidence
in this regard can provide more detailed pictures of how such (counter-attitudinal) messages are
actually selected, processed, and adopted by an individual given exposure settings (e.g.,
Messing & Westwood, 2014; Nyhan et al., 2017; Wood & Porter, 2018). Yet as previously
suggested, a typical experimental approach tends to be misspecified in terms of crucial social
dynamics surrounding simple vs. complex contagion of misinformation and correction
messages within a naturally-occurring social network (Centola, 2010; Margolin et al., 2017).
Moreover, designing a realistic experiment involving real-world social interactions often
involves significant practical (e.g., Bond et al., 2012) and ethical challenges (e.g., Kramer,
Guillory, & Hancock, 2014) for researchers.

Similarly to attitudinal composition of citizen's social networks, the question of how
exactly the structure of citizens' social networks exhibits certain topological properties have
attracted considerable interests among scholars. Prior observations on this topic suggests that
(online) social networks tend to exhibit "small-world" like properties (Kumar, Novak, &
Tomkins, 2010; Ugander, Karrer, Backstrom, & Marlow, 2011), where most nodes can be
reached from every other nodes by a small number of steps. This is due to the fact that a small
minority of "hub" nodes posses a disproportionate number of links, providing a global bridge
between smaller, more strongly interconnected local clusters (Barabási, 2004). This structure
has traditionally been known for information to be spread globally through such hubs (e.g.,
Bakshy et al., 2012), following Granovetter's (1977) seminal account of strength of weak ties.
However, emerging evidence suggests that such a small-world structure may hinder a given
diffusion process to globally saturate the network, since such hubs can serve as a "bottleneck"
of a contagion process, therefore diffusions are stay localized and compartmentalized (e.g.,
Centola & Macy, 2007; González-Bailón, 2017; Zhao, Wu, & Xu, 2010). Similarly, while the
existence of tightly-knit neighborhoods in a typical small-world network supports a
complex-type behavioral propagation (i.e., an innovation that runs counter to prevalent norms

and values: Centola, 2010), the existence of hubs may not favor complex contagion to be globally propagated into a given network. This is due to the fact that such hubs require a much higher number of active neighbors to meet the threshold of adopting a controversial innovation (i.e., complex contagions), and even if such hubs are activated, they alone cannot further activate their immediate neighbors to adopt such innovations in the absence of other local nodes that provide multiple independent reinforcements (Centola, Eguíluz, & Macy, 2007; González-Bailón, 2017; Zhao et al., 2010).

Therefore, whether topological properties of a given network actually hinder or facilitate a contagion process may critically dependent upon specific thresholds values for complex contagions and its interaction with the structure of network (e.g., Centola et al., 2007), which is often hard to be tractable empirically. Moreover, especially for the case of misinformation and its corrections, a simple contagion and a complex contagion may simultaneously affect a local-level attitudinal composition of one's neighbors (i.e., those who endorse and believe misinformation vs. those who do not: e.g., Campbell & Salathé, 2013). Last but not least, all of aforementioned factors are endogenously determined over time in evolving dynamic network. As a consequence, a proper identification regarding the exact impacts of its attitudinal composition and structural properties of a given network often necessitates systematic comparisons between the observed network and a number of plausible counterfactuals. Doing so further requires (a) the ability to compared different thresholds values for complex contagions, and (b) the ability to independently manipulate topological structures of a social network (e.g., Centola, 2010; González-Bailón, 2017). For this reason, a rigorous empirical test of such factors is often impossible for observational studies despite its scientific and practical importance.

A Current Investigation: Mathematical Modeling of Complex Diffusion Dynamics

Adopting a well-known class of an epidemic model of virus infection and recovery, this study used stochastic network-based mathematical models to simulate contagion dynamics of misinformation and its corrections. More detailed methods, including model parameterization, simulation, and more detailed analyses of simulation model results, are provided in the supplementary appendix. Behavioral parameters governing the network structure and misinformation propagation dynamics were estimated from several best-available existing

nationwide representative surveys concerning citizen's social networks (American National

Election Study 2008–2009 Panel Survey: ANES, 2009) and exposure to fake news during 2016

presidential election (Allcott & Gentzkow, 2017; Guess et al., 2018; Pew Research Center,

2016). Although these data sources were not specifically designed for obtaining the impact of

(mis)information diffusion dynamics within a social network, it nevertheless provides a

reasonable starting point for formally modeling more realistic network structures, on which we

later further parameterizing the assumptions and mechanisms postulated in the previous

discussions of complex diffusion dynamics.[4] All simulation models were based on the

EpiModel R software package (Jenness, Goodreau, & Morris, 2018).

Utilizing simulation-based approach in modeling (mis)information propagation is not

entirely new (e.g., Jin, Dougherty, Saraf, Cao, & Ramakrishnan, 2013; Tambuscio, Ruffo,

Flammini, & Menczer, 2015; Zhao et al., 2010). Yet these studies have typically relied on a

series of deterministic compartmental models, where diffusion dynamics were derived by

differential equations representing analytic epidemic systems (e.g., see Zhao et al., 2010).

While they offer valuable insights of how information spread in a given system, they do not

explicitly represent contact phenomena in a network, therefore somewhat limited in realistically

representing the evolving dynamic networks on which information and behaviors are spread

conditional on the network structure. Our approach was to use an exponential random graph

model (ERGM) framework (Morris, Handcock, & Hunter, 2008; Robins, Pattison, Kalish, &

Lusher, 2007) to simulate a series of dynamically evolving networks of different topological

properties (see also below section *A Stochastic Dynamic Model of Networks* for a more detail).

The ERGM framework is now regarded the most versatile yet flexible method for identifying

how a given network is formed and evolves based on the underlying generative principles

(Cranmer, Leifeld, McClurg, & Rolfe, 2017; Robins et al., 2007), and for a simulation context,

it offers a convenient yet powerful tool for independently manipulating topological structures of

a evolving network based on such generative mechanisms by a researcher's choice (Leifeld,

Cranmer, & Desmarais, 2017; Morris et al., 2008). In our present application, we simulated a

baseline network evolution model based on available empirical evidence along with several

counterfactual models that differ in their structural characteristics, and further submit them to

different stochastic models of misinformation propagations vs. corrective message adoptions. Our focus here is therefore on the role of varying network topologies and resulting level of attitudinal homophily as a driving force of network evolutions, and how attitudinal compositions of one's local networks (as end-results of attitudinal homophily) interactively influence simple vs. complex contagion dynamics at population level in conjunction with network topologies.

Therefore, our approach differs from prior studies that have used similar simulation-based approaches in three critical aspects: (a) instead of relying on purely hypothetical scenarios and parameter values, we augment our model by supplying more realistic parameter values, where available, based on existing empirical evidence; (b) our model explicitly takes into account the stochastic nature of network evolution and its endogenous influence over time; and finally, (c) we explicitly model how attitudinal compositions of one's immediate neighbors and local-level contagion dynamics (i.e., simple vs. complex contagions) interact to yield propagation dynamics at the population level. Arguably, this approach obviously still greatly simplifies the complexity of the real world. However, in doing so, it provides an analytical tool for identifying the effects topologies and attitudinal compositions of a network, and the resulting implications for the simple vs. complex diffusion dynamics of misinformation and corrections in a typical social network.

A Stochastic Dynamic Model of Networks

In order to explore role of network topologies in diffusion dynamic in evolving networks, we require to explicitly specify a generative principle of network structure and how such structure evolves over time. To this end, we first model formation and dissolution of citizens' political discussion ties over time through which misinformation and corrections are communicated. Using separable temporal ERGMs (Krivitsky & Morris, 2017), we estimate and simulate *complete networks* that consistently reproduce observed properties of "summary statistics" drawn from egocentric network from 2009 ANES data.[5] Those observed properties, or *summary target statistics*, include: density (i.e., average number of discussion partners, or "alters"), average expected number of degrees among Democrats and Republicans (i.e., main effect of partisanship, excluding Independents), the extent of partisan homophily (i.e., interaction between ego and alter partisanship), and number of people who posses multiple

contacts (i.e., degree distributions).[6] Based on those information, we establish our (a) baseline small-world with political homophily network model, consist of 2400 nodes that comprised of 1258 Democrats and 1142 Republicans, that closely approximate the properties of social networks documented in ANES data. In this model, we set approximately 70% of all exiting ties to be politically homophilous based on observed level of political homogeneity in ANES egocentric networks. We also specify a differential relationship dissolution conditional on partisan homophily of a tie (e.g., selective "unfriending": Noel & Nyhan, 2011), such that any tie within the same partisan group is expected to last longer approximately twice than cross-partisan discussion ties. In addition, in order to facilitate a robust comparison of different network typologies, we additionally specify several competing models of network structure: (b) Erdős–Rényi random network conditional on observed density, (c) a chain network where all nodes have no more than two connections at the same time,[7] and (d) a small-world network *absent of partisan homophily* in tie formation and tie dissolution model. Below Figure 1 shows the simulated cross-sectional views of networks, and Table 1 reports key descriptive statistics for each of the simulated networks from each network evolution model specification in terms of their topological properties.

– Table 1 and Figure 1 about here –

Adapting a well-known class of epidemic models, we simulated the dynamics of misinformation exposure and its correction on top of those stochastic network evolution models. We setup an environment consist of 2400 partisan actors (based on 2009 ANES panel survey's partisan composition: see footnote 6 for a detail) and simulate the diffusion dynamics over 1000 simulated hours (roughly corresponds to 41 days) with 100 replicated simulations. We fit a series of SEIR ("susceptible - exposed - infected - recovered") dynamic models for (a) a simple contagion of misinformation (i.e., a progression from "suspected" to "exposed"), (b) a progression from "exposed" to "infected" (i.e., those who believe a given misinformation conditional on exposure), and (c) a progression from "infected" to "recovered" conditional on exposure and adoption of corrective messages from their peers.

Misinformation Transmission

To date, how many times people exactly "share" fake news and misinformation on their social networks is not yet known to public. However, a recent study by Allcott and Gentzkow (2017) report that, for all 65 fake news websites they have identified during 2016 election, an individual visited such sites 0.64 times on average, and among them 41.8% of all visits originate from social media traffics. Similarly, Guess et al. (2018) report that of 27.4% individuals who visited fake news websites, 22.1% of all visits originate from social media.

We start by setting our baseline prevalence, of which we arbitrary set 15% of partisans are initially affected by misinformation and fake news. We then set our baseline "act rate" (i.e., the average number of transmissible acts per unit time per discussion tie) by taking the average of estimates from Allcott and Gentzkow (2017) and from Guess et al. (2018), which yields 0.164 times per hour per discussion tie.[8] We further assume 15% of those who initially affected are only Republicans, given that the vast majority of identified fake news articles during 2016 U.S. presidential election were overwhelmingly pro-Trump (e.g., Guess et al., 2018). Here, we assume misinformation always flows from the "infected" (i.e., those who already believe given misinformation) to their susceptible peers (i.e., those who are not yet exposed to misinformation), but those who are being merely exposed to do not contribute to misinformation flow and propagation (since the mere exposure does not necessarily mean they are endorsing such information) nor any other sources of exposure to misinformation is possible (i.e., partisan media exposure) for an individual to encounter such misinformation.

Progression from *being susceptible* (i.e., not yet being exposed) to *being exposed* was determined by a simple contagion, such that a single contact with those who are infected (= threshold value of 1) is sufficient for partisans to be exposed to misinformation, yet conditional on the propensity for those who are infected to actually share such news (i.e., the probability of infection: see below). Following PEW data (Pew Research Center, 2016) which suggests approximately 16% of all partisans have had *shared* fake news stories on their social network, we assume the risk of transmission of those who are already infected to share misinformation (i.e., infection probability per act) to be .16 in our simulations. Combining those two parameters for the probability of infection per act (= .16) and the number of acts per unit time (= .164), the

per-discussion tie transmission rate is calculated as an exponential function of the two, with a Bernoulli probability distribution controlling whether an infection event occurs stochastically. In addition, we additionally assume Republicans on average have a higher infection probability per act than Democrats (=.04 higher than Democrats) – once infected, Republicans are more likely to share such news than Democrats – given the directional nature of misinformation being considered within this context (e.g., the vast majority of fake news articles during 2016 U.S. presidential election were overwhelmingly pro-Republican).

Progression from Exposed to Infectious States

Progression from *being exposed* to *being infected* (i.e., transition between exposed to infectious states) was modeled as individual-level stochastic processes, in which there is a differential hazard of transition across partisans following a Bernoulli distribution with the transition rate parameter. Consistent with partisan motivated reasoning accounts (Nyhan & Reifler, 2010; Taber & Lodge, 2006), we set Republicans to have higher transition rates overall (= .30) than Democrats (= .07) given pro-republican misinformation exposure when they are at their mean level of accuracy motivations, where accuracy motivations are assumed to be normally distributed among population. We set these rate parameters based on results reported in Garrett et al. (2016), where they show that online partisan news use increases the chance of incorrectly answering the given factual knowledge question by approximately 30% points for Republicans and 7% point for Democrats compared to those who do not exposed to such information (see their Table S6 for a detail). Also, we assume those who score high in their accuracy motivations to have a lower transition rate regardless of their partisanship (see the online appendix for a detail), effectively treating one's partisanship and accuracy motivations as individual-level moderators of misinformation adoption process.

Corrective Message Exposure and a Recovery from Misperceptions

For corrective message exposure and subsequent progression from *being infected* to *being recovered*, we examine two possible scenarios – the first model closely approximate a single-shot, *asocial* correction, in that the model assumes corrections are randomly received for infected individuals, following individual-level stochastic process of adopting such corrections given exposure. This model therefore assumes no social-dynamics surrounding the

misinformation correction, as typically assumed in extant experimental literature. Our second, more realistic model is more consistent with our prior discussion of simple vs. complex contagion dynamics, in that once exposed, individuals progress to recovery state when a certain proportion (= ρ) of his or her neighbors are sending corrective messages that debunking misinformation. In these scenarios, we assume approximately half of one's immediate neighbors ("alters") who are not yet affected and/or already being recovered from misinformation send correction messages to the infected ego. We systematically vary the threshold value of ρ to explore system-level consequences of social dynamics surrounding misinformation corrections. Below Table 2 lists all scenarios for the combination of network topological structure and the threshold value of ρ in misperception recovery.

– Table 2 about here –

Results

Simple Contagions of Misinformation Exposure

We first look at the simple contagion dynamics of misinformation exposure (assuming no recovery by corrections throughout the simulated time periods) and propagation rates across different network typologies over time, as depicted in Figure 2 and in Table 3 below. The results reveal that network structures have somewhat non-trivial effects on the dynamics of misinformation propagation at population level. It seems that our baseline network model (i.e., small-world network with partisan homophily) appears to be bit better in terms of disseminating given misinformation more faster than other network structures – that is, we do find some modest evidence that spread of simple misinformation spreads faster in our small world networks (red lines in Panel A of Figure 2) than in random networks (black lines in Panel A of Figure 2) when misinformation is spread via simple contagion (i.e., a single contact with those who are infected) followed by a individual-level stochastic progression to being infected given exposure to misinformation.

– Table 3 and Figure 2 about here –

Although differences among the cumulative exposure fractions in different network structures were not appear to be vastly different from each other (as their confidence intervals of cumulative exposure fractions significantly overlap with each other: see online supplementary

materials for a detail), their exposure rates (as measured by the number of newly exposed individuals at given time point) do differ with each other as can be seen in Figure 2 below. This was especially true between our baseline network (red lines) and random networks (black lines) as indicated in Table 4. We used the Wilcoxon Rank Sum Test (a nonparametric test of differences of two population distributions) to test whether the two proportions of newly exposed to misinformation among the susceptible (= no. of SI flow divided by no. of susceptible per each time point) significantly differ from each other between different network topologies. The cell entries of Table 4 refer to the range of simulated time points in which two distributions of adoption rates statistically different from each other based on Wilcoxon Rank Sum Test. For instance, Table 4 indicates that from 58 to 901 time points, the distribution of the fraction of newly exposed to misinformation from ER random network and that from our baseline small-world homophily network statistically different from zero (less than .05). The results shows that across range of time points excluding the very initial stage of misinformation propagation, network structure induces statistically different adoption rates over time.

– Table 4 about here –

Complex Contagions of Misinformation Corrections

We now examine complex contagion dynamics of corrective message adoption and its impacts on population-level dynamics of misinformation propagation (detailed recovery scenarios are described in Table 2). Table 5 lists results for all scenarios in terms of prevalence of misinformation and number of those who are newly being recovered per hour, and Panel A to Panel D in Figure 3 graphically depicts these results.

– Table 5 and Figure 3 about here –

As can be seen in Figure 3, all of the social-based correction models produced more effective results in deterring misinformation propagation in the onset of correction intervention. As time progresses, however, across all of the network topologies we instead see asocial random correction models tend to produce more number of being recovered from misinformation infection. However, this is due to the fact that in such scenarios, the number of those who are being *infected* are also larger than threshold-based correction scenarios – since the number of being recovered is dependent upon the number of being infected at previous time point, even

seemingly effective correction settings may produce much fewer cases of being recovered if it had averted new infection at the first place. Therefore, instead of looking at cumulative mean *prevalences* over time, Figure 4 below alternatively present Proportions of infections being averted (PIA) relative to no-recovery scenarios reported in Table 3. For this metric, we derive the proportion of those who otherwise being infected if no recovery had been taken place at each simulated time point as follow:

$$PIA_t = \frac{I_t^{baseline} - I_t^{recover}}{I_t^{baseline}}$$

where I is the number of those who are infected in each of the recovery scenarios $I^{recover}$ (as in Table 2) vs. no-recovery scenarios $I^{baseline}$ (as in Table 3) at each simulated hours t. Since no-recovery scenarios reported in Table 3 are based on the same network typologies and model specifications *except* infection recovery settings, this allow us to empirically identify counterfactuals without hypothetically assuming number of those who stay being infected. We first derive PIA values for each simulated time points (from time 1 to time 1000), and then take the mean across all time points, then replicate this estimate across 100 simulations per model, as plotted in Figure 4.

As can be seen in Figure 4, our results reveal that in all of the network topologies, socially-contingent corrections produce much more effective intervention results than compared to asocial corrections typically considered in extant literature. All of the asocial correction scenarios yield on average PIA values of .46 to .47, whereas all thresholds-based models typically yield PIA values of .74 to .80, meaning on average across all time points, 74 to 80% of those who otherwise being stay infected and/or newly infected can be prevented based on social-based correction messages exposure and complex contagion of such messages. This appears to be the function of the fact that, in earlier onset of correction message exposure, thresholds values required to adopt correction messages are likely to be easily met due to the scarcity of misinformation in one's immediate environments, therefore attitudinal compositions surrounding simple contagion of misinformation vs. complex contagion of corrective messages tends to favor complex contagion dynamics. This is also evident in Figure 4, where models with much higher threshold value (=.75, in recovery model 4, model 8, model 12 and model 16) are

slightly less effective than other threshold-based models. However, it is indeed interesting to observe that even relatively low threshold (=.25) and moderate threshold (=.50) values all social correction models greatly enhance the effectiveness of corrective message exposure and their adoption.

Sensitivity Analysis

Since we simply assumed the probability of alters sending corrections messages to an infected ego to be .5 in our main results, a series of sensitivity analysis were conducted in order to evaluate the influence of this parameter on the level of PIA metric. In the analyses, probabilities of alters sending corrections to an infected ego were varied to be 0.25 vs. 0.5 vs. 0.75. Figure 5 below reports the results of this sensitivity analysis. Both factors appear to have some impact on PIA estimates, yet at the low to average thresholds models (i.e., models that assume thresholds of adopting corrective messages to be 0.25 or 0.5), the probability of alters sending correction messages did not greatly alter the reported findings in Figure 4. In the high thresholds models (=0.75, blue plots in Figure 5), we see the the increased thresholds specification marginally decreases the PIA, yet the higher value of probability of sending corrections appeared to mitigate the effect of higher threshold parameters, especially in our baseline small-world homophily network models (model 14 to model 16 in Table 2). In all of the models, however, PIA metrics are still substantially higher than single-shot, asocial correction scenarios (typically from .46 to .47).

<div align="center">

Discussion

</div>

TBA

References

Allcott, H. & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*, 211–36.

ANES. (2009). The american national election studies 2008-2009 panel study [dataset]. Stanford University and the University of Michigan [producers and distributors].

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, *348*(6239), 1130–1132.

Bakshy, E., Rosenn, I., Marlow, C., & Adamic, L. (2012). The role of social networks in information diffusion. In *Proceedings of the 21st international conference on world wide web* (pp. 519–528). WWW '12. Lyon, France: ACM.

Barabási, A. L. (2004). *Linked: How everything is connected to everything else and what it means for business, science, and everyday life*. New York: Plume.

Bode, L. & Vraga, E. K. (2017). See something, say something: Correction of global health misinformation on social media. *Health communication*, *Advanced online publication*, 1–10.

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature*, *489*(7415), 295.

Campbell, E. & Salathé, M. (2013). Complex social contagion makes networks more vulnerable to disease outbreaks. *Scientific Reports*, *3*, 1905.

Centola, D. (2010). The spread of behavior in an online social network experiment. *Science*, *329*(5996), 1194–1197.

Centola, D., Eguíluz, V. M., & Macy, M. W. (2007). Cascade dynamics of complex propagation. *Physica A: Statistical Mechanics and its Applications*, *374*(1), 449–456.

Centola, D. & Macy, M. (2007). Complex contagions and the weakness of long ties. *American Journal of Sociology*, *113*, 702–734.

Chan, M. S., Jones, C. R., Jamieson, K. H., & Albarracín, D. (2017). Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological Science*, *28*, 1531–1546.

Cranmer, S. J., Leifeld, P., McClurg, S. D., & Rolfe, M. (2017). Navigating the range of statistical tools for inferential network analysis. *American Journal of Political Science*, *61*(1), 237–251.

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., . . . Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, *113*(3), 554–559.

Delli Carpini, M. X. & Keeter, S. (1996). *What americans know about politics and why it matters*. New Haven, CT: Yale University Press.

Egelhofer, J. L. & Lecheler, S. (2017, September). *Conceptualizing "fake news" for political communication Research: A Framework and Research Agenda*. Paper presented at The Third Annual IJPP Conference, Oxford, UK.

Flynn, D., Nyhan, B., & Reifler, J. (2017). The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics. *Political Psychology*, *38*(S1), 127–150.

Friedkin, N. E. (2001). Norm formation in social influence networks. *Social Networks*, *23*, 167–189.

Garrett, R. K. (2017). The "echo chamber" distraction: Disinformation campaigns are the problem, not audience fragmentation. *Journal of Applied Research in Memory and Cognition*, *6*, 370–376.

Garrett, R. K., Nisbet, E. C., & Lynch, E. K. (2013). Undermining the corrective effects of media-based political fact checking? The role of contextual cues and naïve theory. *Journal of Communication*, *63*, 617–637.

Garrett, R. K., Weeks, B. E., & Neo, R. L. (2016). Driving a wedge between evidence and beliefs: How online ideological news exposure promotes political misperceptions. *Journal of Computer-Mediated Communication*, *21*, 331–348.

Giglietto, F., Iannelli, L., Rossi, L., & Valeriani, A. (2016). Fakes, news and the election: A new taxonomy for the study of misleading information within the hybrid media system. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2878774

González-Bailón, S. (2017). *Decoding the social world: Data science and the unintended consequences of communication*. Boston, MA: MIT Press.

Granovetter, M. S. (1977). The strength of weak ties. *Social Networks*, 347–367.

Guess, A., Nyhan, B., & Reifler, J. (2018). Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 us presidential campaign. Retrieved from https://www.dartmouth.edu/~nyhan/fake-news-2016.pdf

Huckfeldt, R., Mendez, J. M., & Osborn, T. (2004). Disagreement, ambivalence, and engagement: The political consequences of heterogeneous networks. *Political Psychology*, *25*(1), 65–95.

Jenness, S., Goodreau, S. M., & Morris, M. (2018). *Epimodel: Mathematical modeling of infectious disease dynamics*. R package version 1.6.1, http://cran.r-project.org/package=EpiModel.

Jin, F., Dougherty, E., Saraf, P., Cao, Y., & Ramakrishnan, N. (2013). Epidemiological modeling of news and rumors on twitter. In *Proceedings of the 7th workshop on social network mining and analysis* (p. 8). ACM.

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, *111*(24), 8788–8790.

Krivitsky, P. N. & Morris, M. (2017). Inference for social network models from egocentrically sampled data, with application to understanding persistent racial disparities in hiv prevalence in the us. *The annals of applied statistics*, *11*(1), 427.

Kuklinski, J. H., Quirk, P. J., Jerit, J., Schwieder, D., & Rich, R. F. (2000). Misinformation and the currency of democratic citizenship. *The Journal of Politics*, *62*(3), 790–816.

Kumar, R., Novak, J., & Tomkins, A. (2010). Structure and evolution of online social networks. In *Link mining: Models, algorithms, and applications* (pp. 337–357). Springer.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*, 480–498.

Larson, J., Nagler, J., Ronen, J., & Tucker, J. (2016, June). *Social networks and protest participation: Evidence from 93 million twitter users*. Paper presented at the 9th Annual

Political Networks Workshops & Conference, Washington University in Saint Louis, MO. Retrieved from https://ssrn.com/abstract=2796391

Lazer, D., Baum, M., Grinberg, N., Friedland, L., Joseph, K., Hobbs, W., & Mattsson, C. (2017). Combating fake news: An agenda for research and action. *Harvard Kennedy School, Shorenstein Center on Media, Politics and Public Policy*, *2*.

Lazer, D., Rubineau, B., Chetkovich, C., Katz, N., & Neblo, M. (2010). The coevolution of networks and political attitudes. *Political Communication*, *27*, 248–274.

Lee, B. & Bearman, P. (2017). Important matters in political context. *Sociological Science*, *4*, 1–30.

Lee, J. & Song, H. (2017). Why people post news on social networking sites: A focus on technology adoption, media bias, and partisanship strength. *Electronic News*, *11*, 59–79.

Leifeld, P., Cranmer, S. J., & Desmarais, B. A. (2017). Temporal exponential random graph models with btergm: Estimation and bootstrap confidence intervals. *Journal of Statistical Software*.

Levitan, L. C. (2017). Social constraint and self-doubt: Mechanisms of social network influence on resistance to persuasion. *Political Psychology*, *Advanced online publication*.

Levitan, L. C. & Visser, P. S. (2008). The impact of the social context on resistance to persuasion: Effortful versus effortless responses to counter-attitudinal information. *Journal of Experimental Social Psychology*, *44*, 640–649.

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, *13*, 106–131.

Lewandowsky, S., Ecker, U. K., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the "Post-Truth" era. *Journal of Applied Research in Memory and Cognition*, *6*(4), 353–369.

Margolin, D. B., Hannak, A., & Weber, I. (2017). Political fact-checking on twitter: When do corrections have an effect? *Political Communication*, *Advanced online publication*, 1–24.

Messing, S. & Westwood, S. J. (2014). Selective exposure in the age of social media: Endorsements trump partisan source affiliation when selecting news online. *Communication Research*, *41*, 1042–1063.

Metzger, M. J., Flanagin, A. J., & Medders, R. B. (2010). Social and heuristic approaches to credibility evaluation online. *Journal of communication*, *60*(3), 413–439.

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*, 1–12. Retrieved from https://doi.org/10.1371/journal.pone.0184148

Morey, A. C., Eveland, W. P., Jr., & Hutchens, M. J. (2012). The "who" matters: Types of interpersonal relationships and avoidance of political disagreement. *Political Communication*, *29*(1), 86–103.

Morris, M., Handcock, M. S., & Hunter, D. R. (2008). Specification of exponential-family random graph models: terms and computational aspects. *Journal of statistical software*, *24*(4), 1548.

Noel, H. & Nyhan, B. (2011). The "unfriending" problem: the consequences of homophily in friendship retention for causal estimates of social influence. *Social Networks*, *33*(3), 211–218.

Nyhan, B., Porter, E., Reifler, J., & Wood, T. (2017). Taking corrections literally but not seriously? The effects of information on factual beliefs and candidate favorability. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2995128

Nyhan, B. & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, *32*, 303–330.

Pew Research Center. (2016, December). Many americans believe fake news is sowing confusion. Retrieved from http://www.journalism.org/2016/12/15/many-americans-believe-fake-news-is-sowing-confusion/

Robins, G., Pattison, P., Kalish, Y., & Lusher, D. (2007). An introduction to exponential random graph (p*) models for social networks. *Social networks*, *29*(2), 173–191.

Shearer, E. & Gottfried, J. (2017, September). News use across social media platforms 2017. Retrieved from

http://www.journalism.org/2017/09/07/news-use-across-social-media-platforms-2017/

Siegel, D. A. (2009). Social networks and collective action. *American Journal of Political Science*, *53*(1), 122–138.

Small, M. L., Pamphile, V. D., & McMahan, P. (2015). How stable is the core discussion network? *Social Networks*, *40*, 90–102.

Song, H. (2015). Uncovering the structural underpinnings of political discussion networks: Evidence from an exponential random graph model. *Journal of Communication*, *65*, 146–169.

Taber, C. S. & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, *50*, 755–769.

Tambuscio, M., Ruffo, G., Flammini, A., & Menczer, F. (2015). Fact-checking effect on viral hoaxes: a model of misinformation spread in social networks. In *Proceedings of the 24th international conference on world wide web* (pp. 977–982). ACM.

The New York Times. (2017, November). Signs of russian meddling in brexit referendum. Retrieved from https://www.nytimes.com/2017/11/15/world/europe/russia-brexit-twitter-facebook.html?smid=tw-share

Thorson, E. (2016). Belief echoes: The persistent effects of corrected misinformation. *Political Communication*, *33*, 460–480.

Ugander, J., Karrer, B., Backstrom, L., & Marlow, C. (2011). The anatomy of the Facebook social graph. *CoRR*, *abs/1111.4503*. arXiv: 1111.4503

Weeks, B. E. (2015). Emotions, partisanship, and misperceptions: How anger and anxiety moderate the effect of partisan bias on susceptibility to political misinformation. *Journal of Communication*, *65*, 699–719.

Weeks, B. E. & Garrett, R. K. (2014). Electoral consequences of political rumors: Motivated reasoning, candidate rumors, and vote choice during the 2008 us presidential election. *International Journal of Public Opinion Research*, *26*, 401–422.

Wood, T. & Porter, E. (2018). The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior*, 1–29.

Wu, F. & Huberman, B. A. (2007). Novelty and collective attention. *Proceedings of the National Academy of Sciences*, *104*(45), 17599–17601.

Zhao, J., Wu, J., & Xu, K. (2010). Weak ties: subtle role of information diffusion in online social networks. *Physical Review E*, *82*(1), 016105.

Footnotes

[1] Often, the term *fake news* is used as derogatory, rhetorical label to attack political opponents. While such a use of the term as a *label* is an important conceptual dimension to consider, this aspect of *fake news* is beyond the scope of this manuscript. See Egelhofer and Lecheler (2017) instead for a detailed conceptualization involving this distinction.

[2] Here, we use the term "attitudinal composition" to denote a relative distribution of supporters vs. opponents regarding a given attitude or a behavior being spread in a network.

[3] Indeed, in a completely segregated network where cross-ideological links are not present, a simple diffusion of partisan misinformation is not likely to saturate the entire network, since at least some segments of populations are never exposed to such information due to the lack of cross-ideological exposure. In light of our discussion, an "exposure" echo chamber actually *prohibits* global-scale (mis)information propagation.

[4] It should be acknowledged that those sources of data is suboptimal in number of regards: (a) social network data is confined to cross-sectional, egocentric network data, and (b) they do not come from the identical data generating process covering the same periods of observations (e.g., networks recoded in 2009 has little or no bearing for fake news exposure in 2016). However, research indicates the structural compositions of citizens' core discussion network are reasonably stable over time, albeit actual responses in typical surveys are subject to a number of survey-context related factors (Lee & Bearman, 2017) or changes in one's life contexts (Small, Pamphile, & McMahan, 2015). Also, since most of the prior applications concerning fake news and misinformation from observational studies typically look at representative samples, we intentionally have chosen those data since these types of data are more commonly used in the social sciences, and provide best-available up-to-date relevant information for the current application. See also *A Stochastic Dynamic Model of Networks* section below and supplemental appendix for a detail.

[5] Of course, this approach requires certain assumptions in order to define a model that represents a distribution of networks that are centered on the observed properties. See Krivitsky and Morris (2017) for a detailed discussion of this class of models and their properties.

[6] In 2009 ANES panel, there are a total of 1258 self-identified Democrats and 1142 self-identified Republicans (including respective party leaners), excluding 336 Independents. Those 2400 partisans (i.e., Democrats plus Republicans) have on average 1.975 discussion partners. Since in 2009 ANES respondents theoretically name up to three alters, this translate into egonetwork density of 0.66 (=1.975/3). For the main effect of partisanship, the marginal distribution of number of named alters by egos did not differ by partisanship, meaning that the expected mean degree for each partisan group is proportional to observed density and the size of each group. For political homophily, data indicate that 23% of reported discussion ties exist across Democrats and Republicans. For degree distribution statistic, we supply a range of plausible yet arbitrary values since 2009 ANES does not have summary network size measure. See supplemental appendix for a detail.

[7] This is done by first simulating a lattice network (e.g., all nodes have exactly the same number of degrees), but due to the stochastic nature of evolving network in which ties are randomly created and dissolved, the resulting network may not have the uniform degree distribution than it would have been under strict lattice network.

[8] Allcott and Gentzkow report $0.64 \times 41.8\% = 0.268$ shares per unit time, while Guess et al. report $0.274 \times 22.1\% = 0.060$ shares per unit time. This averages to 0.164 times per unit time. We additionally assume all fake news website visits are driven by information sharing on social media.

Table 1

Descriptive Statistics of Cross-sectional Networks from Simulated Network Structures

	E-R	**Chain**	**SW No-Homophily**	**SW Homophily**
Graph density	0.000269	0.000268	0.000277	0.000272
Mean degree	1.292	1.287	1.327	1.303
Max degree	10.0	4.0	12.0	12.0
Clustering coefficients	0.0	0.0	0.02	0.01
Mean distance	2.2	2.2	4.5	3.4
Homophilous ties	49.4%	50.5%	52.3%	68.5%
Heterophilous ties	50.6%	49.5%	47.7%	31.5%

Note: E-R: Erdos Renyi random network. Chain: Chain network. SW No-Homophily: Small-world network absent of homophily. SW Homophily: Small-world network with partisan homophily.

Table 2

Misperception Recovery Scenarios by Network Topologies and Thresholds for Adopting Corrections, Correction Probability = 0.5

	Scenario I	Scenario II - Thresholds		
	Asocial	**0.25**	**0.5**	**0.75**
E-R	Model 1	Model 2	Model 3	Model 4
Chain	Model 5	Model 6	Model 7	Model 8
SW No-Homophily	Model 9	Model 10	Model 11	Model 12
SW Homophily	Model 13	Model 14	Model 15	Model 16

Table 3

*Misinformation Exposure Prevalence Per 100 Individuals Over Time by Network Topologies
(Cumulative Exposure, No Recovery Scenarios)*

|	E-R	Chain	SW No-Homophily	SW Homophily
Overall	.594	.584	.605	.640
(95% CIs)	[.564, .624]	[.558, .609]	[.574, .635]	[.608, .669]
Among Dem	.295	.290	.301	.312
(95% CIs)	[.274, .316]	[.271, .308]	[.279, .322]	[.287, .332]
Among Rep	.299	.295	.305	.330
(95% CIs)	[.285, .313]	[.282, .307]	[.290, .319]	[.315, .343]

Note: E-R: Erdos Renyi random network. Chain: Chain network. SW No-Homophily: Small-world network absent of homophily. SW Homophily: Small-world network with partisan homophily. All statistics are based on 100 random simulations over 1000 time points from the same setup per network topologies.

Table 4

*Time Durations Within Which WRS Test of Differences on Fraction of Newly Exposed of
Misinformation Over Time are Significantly Different From Each Other Across Topologies*

	E-R	Chain	SW No-Homophily	SW Homophily
Chain	106 : 278	-		
SW No-Homophily	68 : 1000	70 : 585	-	
SW Homophily	58 : 901	63 : 942	151 : 1000	-

Note: E-R: Erdos Renyi random network. Chain: Chain network. SW No-Homophily: Small-world network absent of homophily. SW Homophily: Small-world network with partisan homophily. Cell entries denote the range of time points in which Wilcoxon Rank Sum test indicates the adoption rates from two network topologies statistically different from each other ($p < 0.05$).

Table 5

Asocial vs. Thresholds Recovery Model Results

		Scenario I	Scenario II - Thresholds		
		Asocial	**0.25**	**0.5**	**0.75**
E-R	I Prevalence	23.484 [20.663, 26.037]	3.661 [3.070, 4.406]	3.722 [3.031, 4.503]	5.398 [4.383, 6.844]
	R Incidence	1.262 [.000, 3.716]	.429 [.150, 1.350]	.423 [.130, 1.383]	.520 [.0433, 1.944]
Chain	I Prevalence	22.746 [20.051, 25.492]	3.581 [2.983, 4.286]	3.581 [2.983, 4.286]	5.260 [4.259, 6.399]
	R Incidence	1.196 [.000, 3.625]	.405 [0.138, 1.339]	.405 [0.138, 1.339]	.483 [.0372, 1.843]
SW No-Homophily	I Prevalence	24.107 [21.220, 27.047]	3.926 [3.266, 4.693]	4.016 [3.347, 4.821]	6.107 [4.951, 7.639]
	R Incidence	1.329 [.000, 3.797]	.452 [.176, 1.465]	.457 [.150, 1.483]	.599 [.050, 2.150]
SW Homophily	I Prevalence	26.039 [23.029, 29.007]	3.863 [3.148, 4.724]	3.983 [3.295, 4.884]	6.014 [4.935, 7.423]
	R Incidence	1.420 [.003, 4.003]	.495 [.188, 1.461]	.494 [.168, 1.508]	.656 [.0771, 2.127]

Note: E-R: Erdos Renyi random network. Chain: Chain network. SW No-Homophily: Small-world network absent of homophily. SW Homophily: Small-world network with partisan homophily. I Prevalence: Infection prevalence per 100 individuals over time. R Incidence: Recovery incidence per time.
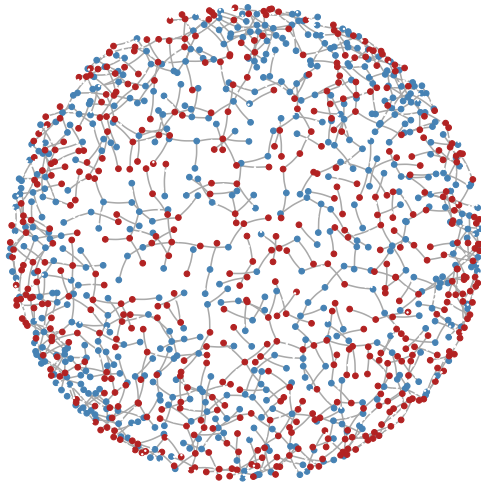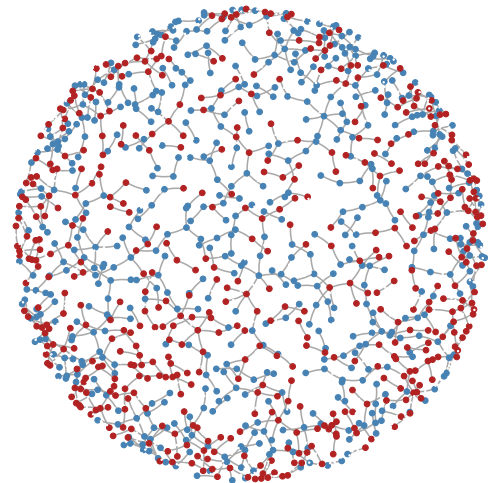
A. Erdos-Renyi random network
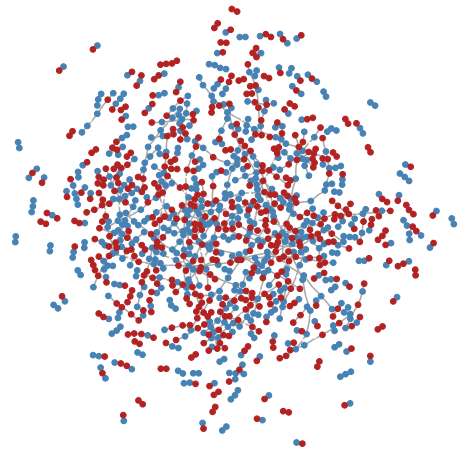
B. Chain network
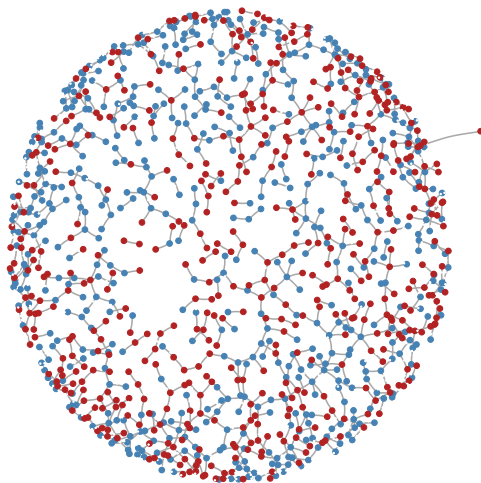
C. Small-world network absent homophily

D. Baseline small-world network

Figure 1. A cross-sectional view of simulated networks from different network topologies.

Note: Colored nodes represent nodes that have at least one connection with their neighbor (Red = Republicans, Blue = Democrats) while isolates at given time point was suppressed in visualization. A Erdos-Renyi (ER) random network (= A) assumes a homogeneous probability of creating and dropping ties across all nodes conditional on observed density. A chain-like network (= B) has more longer chains than the ER random network while lacks triangle-like structure due to constrains on degrees (less than three). While two small-world network has more star-like structure than others, our baseline small-world network (= D) is more homophilous than competing specification in its counterpart without homophily (= C), as can be seen in Table 1.
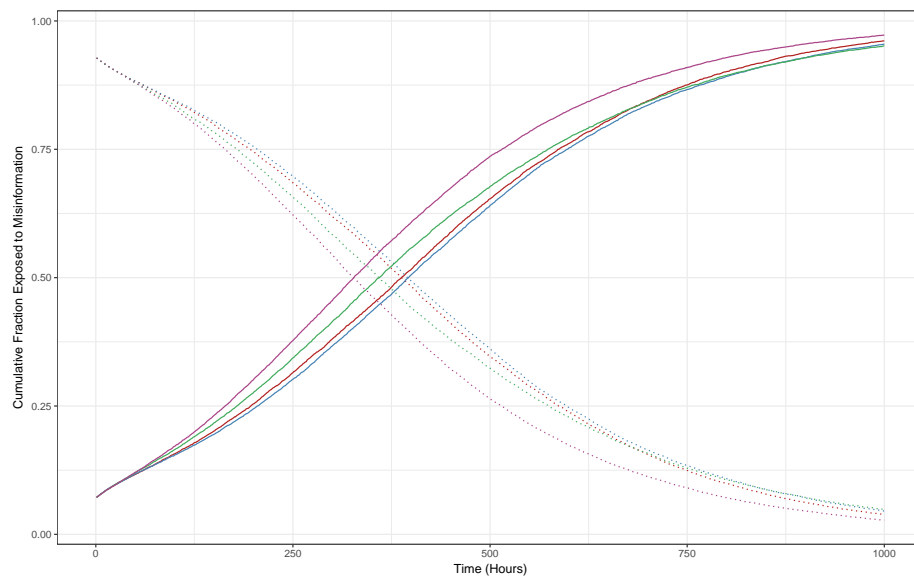
Panel A. Overall Cumulative Prevalence of Misinformation.
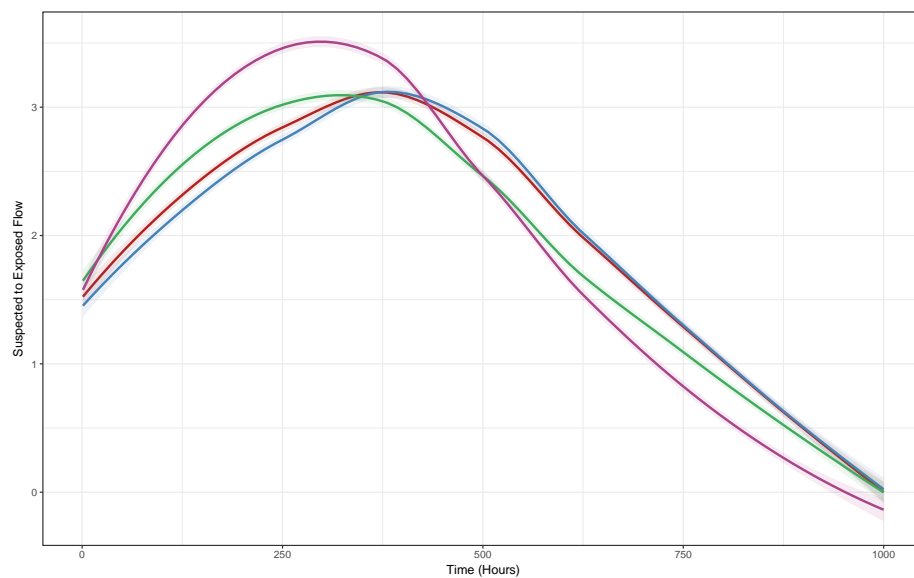
Panel B. Incidence Rates of New Exposure to Misinformation (the number of newly
exposed to misinformation)

Figure 2. Prevalence and Incidence Rates of Misinformation Over Time by Network Topologies
(Assuming No Recoveries, Averaged Over 100 Simulations).

Note: Straight lines in Panel A represent fraction of those who are being exposed to misinformation,
whereas dotted lines represent fraction of those who are susceptible. X-axis represents simulated times
(in hours) and Y-axis represents the fraction exposed to misinformation (in panel A) or rates of
misinformation exposure (in panel B). Red = Erdos-Renyi random network, Blue = Chain network,
Green = Small-world absent homophily, and Violet = baseline small-world homophily network.
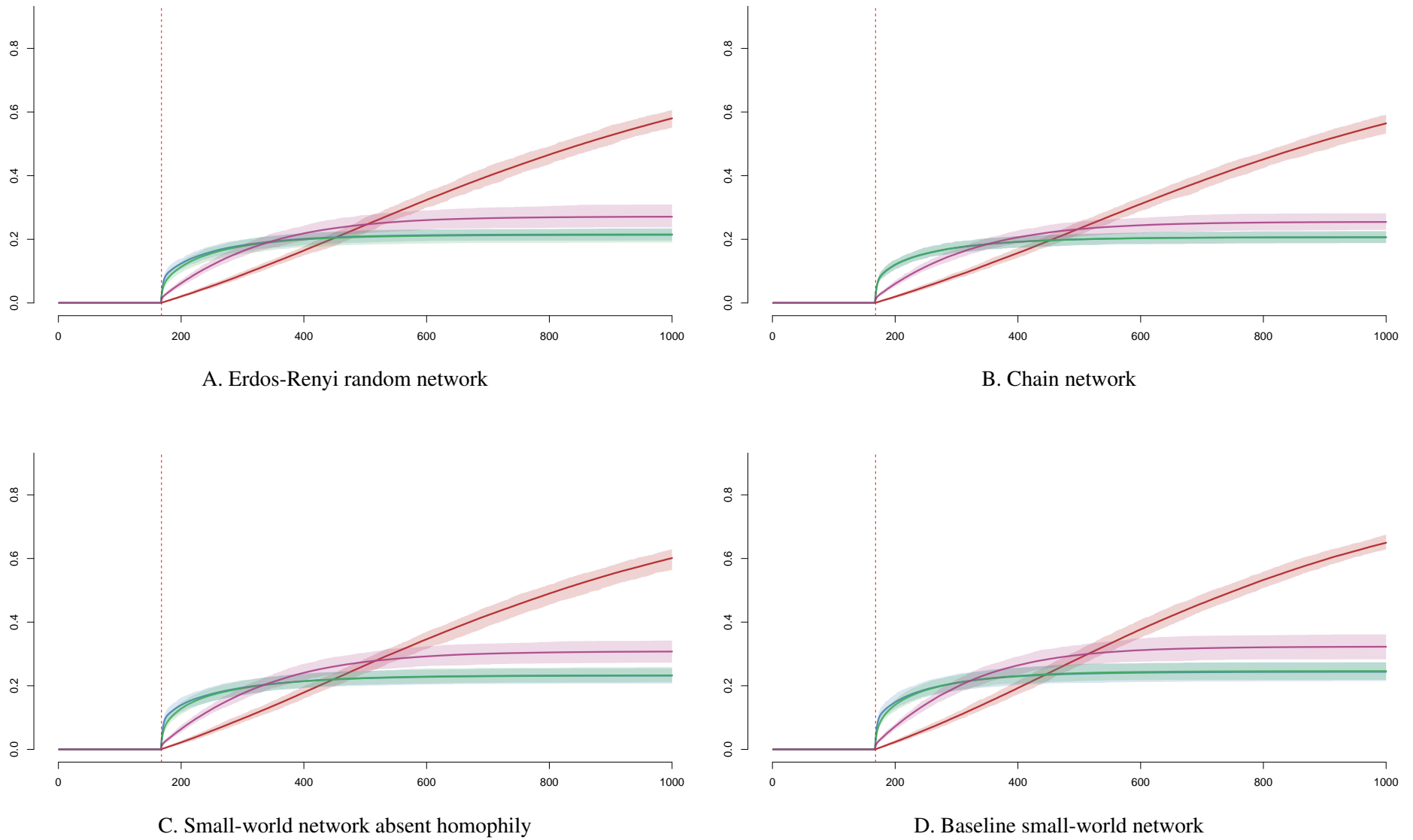
Figure 3. Proportion of Being Recovered from Infection Status, by Network Topologies

Note: Red: Asocial random correction, Blue: Thresholds = 0.25, Green: Thresholds = 0.50, and Violet: Thresholds = 0.75.
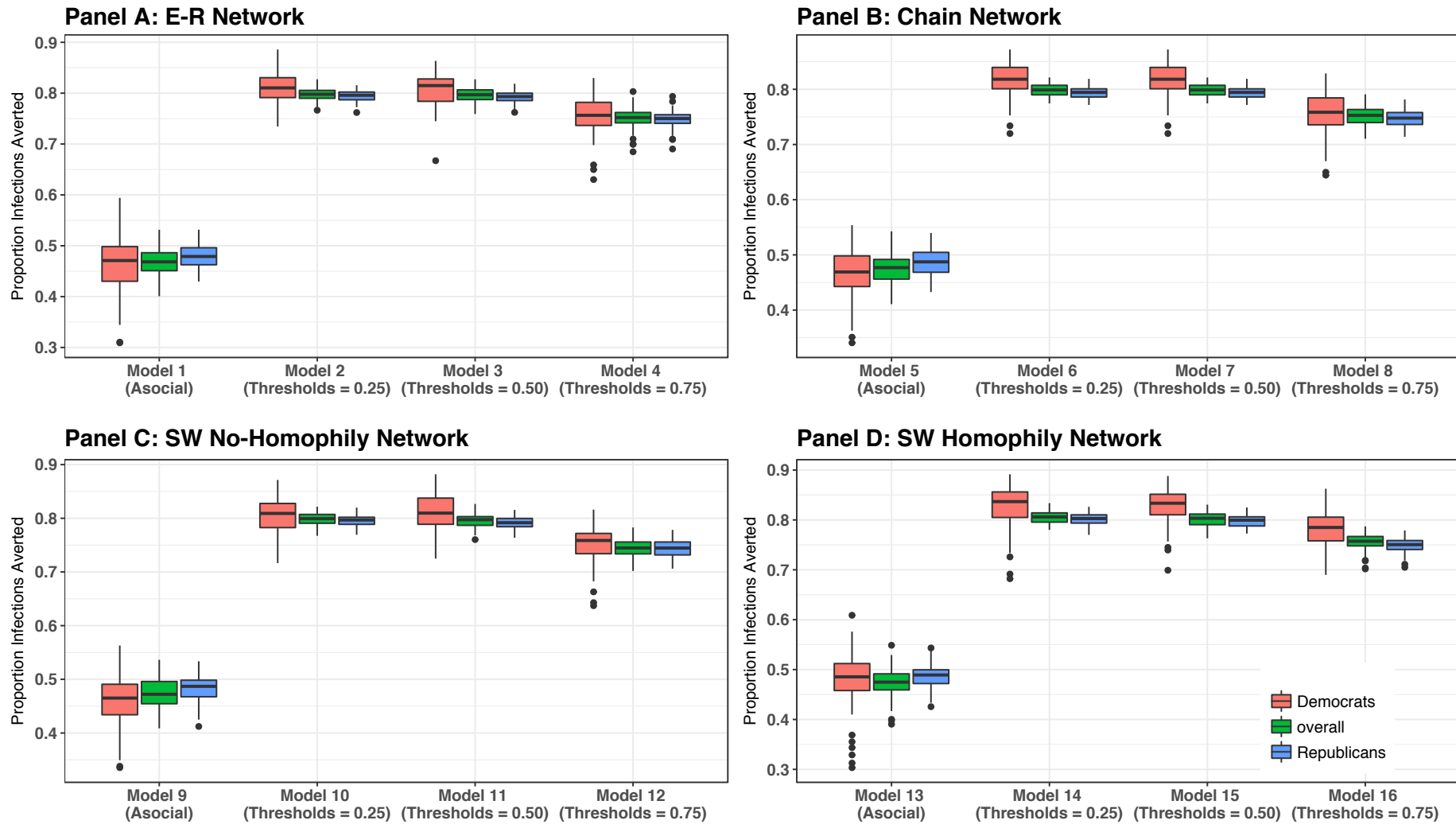
Figure 4. Proportions Infections Averted, Across Recovery Model Thresholds.

Note: Y-axis denotes Proportions of Infections being Averted (PIA) relative to the no-recovery scenarios. We present overall (=green boxplots), among Democrats (=orange boxplots) and among Republicans (=blue boxplots) PIA estimates.
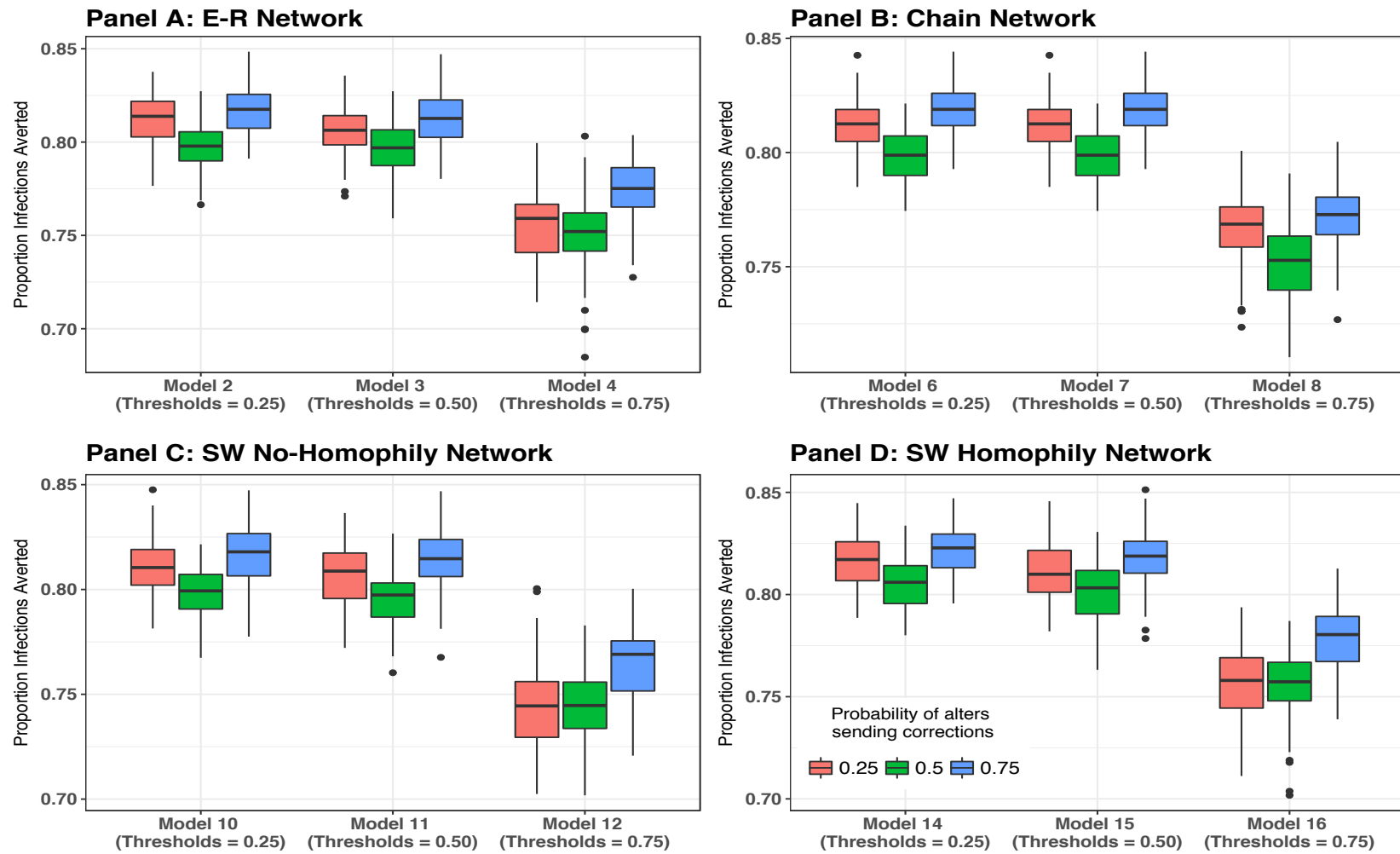
Figure 5. Sensitivity Analysis of Alter Probability of Sending Correction Messages, Across Different Recovery Model Thresholds.

Note: Y-axis denotes Proportions of Infections being Averted (PIA) relative to the no-recovery scenarios. We present different PIA values per each recovery scenario assuming probabilities of alters sending correction messages to infected ego equal to 0.25 (= red boxplots), 0.5 (= green boxplots) and 0.75 (=blue boxplots). Main results reported in 4 assumed probability of alter sending corrections to be 0.5 (= green boxplots).