Proceedings of the 2019 Winter Simulation Conference
N. Mustafee, K.-H.G. Bae, S. Lazarova-Molnar, M. Rabe, C. Szabo, P. Haas, and Y.-J. Son, eds.

AGENT BASED SIMULATION OF BOT DISINFORMATION MANEUVERS IN TWITTER

David M. Beskow
Kathleen M. Carley

School of Computer Science
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA 15213, USA

ABSTRACT

Multiple state and non-state actors have recently used social media to conduct targeted disinformation operations for political effect. Even in the wake of these attacks, researchers struggle to fully understand these operations and more importantly measure their effect. The existing research is complicated by the fact that modeling and measuring a persons beliefs is difficult, and manipulating these beliefs in experimental settings is not morally permissible. Given these constraints, our team designed an Agent Based Model that is designed to allow researchers to explore various disinformation forms of maneuver in a virtual environment. This model mirrors the Twitter Social Media Environment and is grounded in social influence theory. Having built this model, we demonstrate its use in exploring two disinformation forms of maneuver: 1) "backing" key influencers and 2) "bridging" two communities.

1 INTRODUCTION

As more information has become available detailing how Russia's Internet Research Agency (IRA) manipulated various societies and political processes across the world, researchers have worked to document the IRA's methods (Beskow and Carley 2019), develop ways to detect these methods (Beskow and Carley 2018; Chavoshi et al. 2016; Ferrara et al. 2016), and determine how effective these methods are (Bessi and Ferrara 2016).

While these research effort have effectively identified methods, target audiences and developed models to detect certain agents (bots/trolls), it still struggles to answer the question of impact. The primary challenge being that it is difficult to measure and label a person's beliefs, and even harder to measure the evolution of these beliefs over time while enumerating resulting decisions/actions. If is often difficult for a person to identify how a myriad of real and virtual social interactions and messaging affect their own decisions, let alone someone else to measure these. Laboratory experiments that attempt to manipulate a persons beliefs are not morally permissible, therefore human subject experimental studies provide limited utility.

For decades researchers have attempted to fill this gap with models of information diffusion, rumor propagation, and social influence more generally. These models include systems dynamics, discrete event simulation, and more recently agent based models (ABMs). While others have modeled Twitter in these simulations, authors have generally abstracted away the mechanics of the Twitter environment (Tweets, replies, retweets, mentions, following, news feeds, etc). In our effort to model and evaluate disinformation forms of maneuver, we felt that it was important to explicitly model the mechanics of the Twitter environment, and we have not found any research effort that has done this. Our team set out to develop a special purpose ABM called *twitter_sim* where we model a given social media environment (Twitter), insert malicious agents (bots/trolls), conduct various disinformation forms of maneuver, and evaluate the emerging behavior. In

978-1-7281-3283-9/19/$31.00 ©2019 IEEE

addition to exploring disinformation maneuvers, *twitter_sim* could be used for modeling marketing campaigns or as the backbone for a virtual social media training environment.

twitter_sim will include most of the actions of the Twitter environment (tweet, reply, retweet, mention, follow). It will also include heterogeneous human behavior on the platform, including varied rates of access, limited attention, dynamic network links based on homophily and social cues, and changed beliefs based on level of exposure weighted by homophily and authority.

In addition to describing and validating the *twitter_sim* model, we applied it to simulated networks to explore two disinformation forms of maneuver used by bots/trolls: 1) *backing* key influencers and 2) *bridging* two communities. Backing involves following and retweeting influencers to amplify their message. Bridging involves building links between two communities in order to introduce the ideas of one into the other. Our central question is 'What emergent behavior do we observe when bots *back* influencers or *bridge* networks in social media?' These maneuvers were selected because they were used by the IRA in information operations in the United States and Ukraine, and because these methods can be similar.

2 RELATED WORKS

Early models of information diffusion on social media used an epidemiology model known as the SIR Model (Susceptible, Infected, Recovered). Daley and Kendall produced the first model based on this (Daley and Kendall 1965), which was expanded by Maki and Thompson (Maki and Thompson 1973). These models (often referred to as the DK and MK models) describe the three states as ignorants (S), spreaders (I), and stiflers (R). Zanette produced an early model that uses the SIR model in social media (Zanette 2002). There have since been many different evolutions from the original models. Several other works have criticized the epidemiological approach as overly simplistic given it generally assumes a homogenous population connected in a simple network with a constant probability of infection, and because recovery mechanisms of epidemics (often vaccination) are different than the infection mechanisms, whereas in rumors the recovery mechanisms are very similar to the infection mechanisms (i.e. anti-rumor messaging) (Tripathy et al. 2010).

2.1 Agent Based Models of Social Media

Several authors have conducted somewhat similar research using Agent Based Models's (ABM's) to model social media, particularly looking at the spread of rumors or misinformation. Tripathy et al (Tripathy et al. 2010) provides a network based ABM that consists of neutral nodes, infected nodes (believed the rumor), vaccinated nodes (believe the anti-rumor before infection, and cured nodes (believe anti-rumor after infection). Additionally, Tripathy et al explores the idea of a *beacon*, which is similar in purpose to the "stifler" in earlier models. A given authority detects a rumor at some point after the beginning of the rumor (a time delta that they vary and find important) and then position *beacons* which help to broadcast the anti-rumor message. Serrano (Serrano et al. 2015) adapts earlier models by claiming the *recovered* users will not try to influence their neighbors, offering empirical evidence from Twitter to support this. A recent model by Wang et al attempts to capture memory, conformity, differences in propensity to produce/spread rumors, and variance in trust to model rumor propagation with information entropy. (Wang et al. 2017)

2.2 The Construct Model

Some of what we will present in this paper builds on the Construct model originally presented in 1990 by Carley (Carley 1990) and revisited in 2009 (Carley et al. 2009). Construct is a general purpose social influence model that seeks to bring in the complex social dynamics that the are not present in the epidemiological models mentioned above. "Construct is the embodiment of constructuralism, a mega-theory which states that the sociocultural environment is continually being constructed and reconstructed through individual cycles of action, adaptation, and motivation." (Carley et al. 2009) Construct presents the idea that bounded rationality impacts social interaction. Bounded rationality means that agents do not have access to all information (due to social position) and do not process/retain all information that they do have access to. The beliefs of others impact an agent's beliefs through social influence (Friedkin 2006).

Within Construct, the likelihood of interaction is based on relative similarity and relative expertise. Construct agents have general and transactive memory. General memory is the facts it knows and the beliefs it holds. The transactive memory is its view (not necessarily accurate) of "who knows who and who knows what."

While not implementing the full construct model, our model is informed by several concepts that Construct introduces. These include bounded rationality (agents are bounded by position and expertise), likelihood of interaction based on similarity, interactions weighted by similarity and expertise, and a Twitter specific model of general and transactive memory. A docking lite comparison of *twitter_sim* (our model), Construct, and SIR based epidemiological models is provided in Figure 1a.

2.3 The Twitter Environment

Twitter began in 2006 as a way for people to share Short Message Service (SMS) messages with a maximum length of 140 characters. As such, it quickly became the first and arguably the largest of the "micro-blog" platforms. In the Twitter environment, users have a two way following mechanism that is rather unique among social media platforms. Within the Twitter environment, users can interact in a variety of ways. These include the following actions:

1. **Tweet:** Users generate a short message that can include multimedia
2. **Retweet:** Users send another user's message to their followers without comment.
3. **Quote:** Users send another user's tweet to their followers with comment (starting a new thread)
4. **Reply:** A user replies to a tweet that someone else makes (remains in same thread as original tweet)
5. **Mention:** A user places another user's screen name in a tweet; the mentioned user is notified
6. **Like:** Users can like a tweet, which increases its prominence on the platform

Individual's use of Twitter can vary significantly. Some users rarely log on and could be considered dormant. Others use it extremely often, while bots use it at the speed of algorithms. Many models of disinformation on Twitter don't capture this aspect of social media usage, assuming that every user will have a chance of influence in every time step, which doesn't happen in reality.

A Twitter user's feed is only populated by tweets, retweets, and replies produced by those accounts that they follow. While a user can search through the Twitter stream on their own, their feed is only populated by a proprietary algorithm with tweets from those they follow (not those that follow them). *twitter_sim* will explicitly model this structural constraint of the environment.

2.4 Disinformation Maneuvers

The authors have previously outlined the BEND framework for identifying disinformation maneuvers (Beskow and Carley 2019). In the BEND framework, information operations can target both the information and the network. Often, information warfare architects will attempt to manipulate both at the same time. The BEND Framework is summarized in Table 1b.

We developed *twitter_sim* primarily to explore these maneuvers in a virtual environment. From the 16 BEND maneuvers, we selected *back* and *bridge* to initially explore because we hypothesize that *bridging* may simply be the same as *backing* but focused on two communities instead of a single community.

3 MODELING TWITTER ENVIRONMENT WITH AGENT BASED MODEL

We developed *twitter_sim* as an Agent Based Model in Python 3.6. Throughout the development process, we leveraged the rich network functions available in the *networkx* package (Hagberg et al. 2008). Building this in Python on top of *networkx* allows the model to adapt and scale (easily run in parallel on large

Table 1: Tables showing Docking Lite and a summary of BEND framework.

(a) Docking Lite Comparison of *twitter_sim*, Construct, and SIR.

	twitter_sim	Construct	SIR
General Population	✓	✓	✓
Media Agents	✓	✓	
Opinion Leaders	✓	✓	
Information Access	✓	✓	
General Memory	✓	✓	✓
Transactive Memory		✓	
Homophily	✓	✓	
General Populace	✓	✓	
Limited Attention	✓		
Dynamic Network	✓	✓	

(b) Summary of BEND Framework.

	Network	Information
Pro	Back	Engage
	Build	Explain
	Bridge	Excite
	Boost	Enhance
Con	Neutralize	Dismiss
	Nuke	Distort
	Narrow	Dismay
	Neglect	Distract

compute resources), while remaining shareable through open source software mechanisms. Although the models in this paper use scale free networks, the *titter_sim* model was developed to easily accommodate experiments on real world networks and events (i.e. humanitarian disaster or election event).

In *twitter_sim* we explicitly model users and their behavior on Twitter. We use three types of users: normal users (ignorants), bots/trolls (spreaders), and truth defenders (stiflers). Only truth defenders and bots aggressively pursue information operations. Normal users, even once their beliefs begin to change, generally do not aggressively engage in a given campaign (Serrano et al. 2015), but do propagate information messages through retweets.

Normal users and truth defenders begin at time 0 embedded into a preexisting network. Bots start at time zero on the periphery of this network with a single link to a random node. This explicitly models the challenge that bot creators face in embedding and building position in networks and online communities.

As mentioned above, a Twitter users' feed is only populated by content produced by accounts that they follow. It is not populated with the content of people that follow them. *twitter_sim* models this behavior and populates an agent's Twitter feed with tweets from those they follow (their successors). This means that bots, while producing a much higher concentration of disinformation, will only start having an effect once they become embedded in real networks and build a following.

3.1 Twitter use as a discrete event simulation

We have not found a model that captures the fact that Twitter can only influence a user if they log on and read their Twitter feed. The inter-arrival time of people returning to Twitter varies widely. Some people use Twitter multiple times daily, while others check it every other month.

In *twitter_sim* each agent stores the next time step that the agent will log on to Twitter, and won't read their feed, send tweets or adjust their beliefs until that time step. We've modeled the inter-arrival time as an exponential random distribution parameterized by λ, the mean hourly rate. We varied λ with a uniform distribution ranging from 0.001 (once every 2 months) to 0.75 (18 times per day). These numbers are validated with empirical data later in the paper. Therefore inter-arrival time T is defined as

$$T \sim Exponential(\lambda) \quad \text{where} \quad \lambda \sim uniform(0.001, 0.75).$$

3.2 Modeling Limited attention, homophily, and influence

Like other models (Weng et al. 2012), *twitter_sim* will model the limited attention that users have. During a session users will only 'read' the last 4 to 20 tweets in their feed. Only read tweets are used in updating a users beliefs. The number of 'read' tweets is a random uniform integer. This limited attention behavior means that those accounts with a high in-degree will only read a small portion of their total feed, while less popular accounts have the potential to read most of their feed (depending on the rate of activity for them and their followers).

McPhersen et al introduced the idea of homophily in social networks with the idea of "birds of a feather flock together" (McPherson et al. 2001). McPhersen summarized homophily by stating that "similarity breeds connection." *twitter_sim* measures similarity between agents and uses this similarity to create new links as well as weight information (i.e. tweets from agents that are more similar to me will have greater impact on my beliefs).

In our model, homophily, or similarity, is measured by the similarity of out-links (followed accounts or *successors*). If two agents follow many of the same accounts, then they are more similar. The overlap of followed accounts is measured by a Jaccard similarity of the adjacency matrix and is updated on a weekly basis. The Jaccard similarity of User A and User B is therefore computed as follows:

$$similarity = \frac{successors_A \cap successors_B}{successors_A \cup successors_B}.$$

The level of influence for a user is measured as the normalized in-degree of the user. Other models have also used influence to inform probability of acceptance (Wang et al. 2017), though in our case we only use in-degree since Twitter is by nature a directed network.

The number of accounts a user is allowed to follow in Twitter is artificially capped at 5,000 until an account has more than 5,000 followers. In our scale free networks out-degree is limited to a given percentage of the overall nodes in the network (usually 10-20%). In-degree is unconstrained.

3.3 Modeling Mentions, Retweets, *Stiflers*, and the Global Conversation

Twitter allows users to *mention* a user in a tweet. This is done for multiple reasons, and alerts the mentioned user of the tweet. Mentions can also be used by average users as well as bots/trolls in an attempt to gain followers. Our model produces mentions with a given probability, and then a small portion of new links are directed to mentions.

Twitter allows a user to *retweet* another user's tweet. Retweeting without adding additional comment primarily serves the purpose of propagating the message. In our model all agents retweet with a given probability. All retweets carry the value of the original tweet (weighted by the homophily and influence of the originator) as opposed to the retweeter.

On Twitter users actively counter disinformation, and this is often done with *quotes* and *replies*. These users are often referred to as *stiflers* or *beacons*. 10% of the users in *twitter_sim* are labeled as *stiflers*. These users actively combat disinformation by spreading the counter message, which has a negative affect on disinformation belief. *Stiflers* send one reply for every disinformation retweet they read in their inbox. Stiflers are therefore constrained by bounded rationality, and will only counter messages that they are aware of (meaning they are produced by neighbors of the stifler). They are also constrained by limited attention since they only counter tweets they read.

Users are not only influenced by content produced by the accounts they follow, but are also allowed to search through tweets by topic, user, or hashtag. In doing so they are influenced by the larger *global* conversation occurring on Twitter. Our model explicitly models this as the current mean belief of the network and uses it to update a persons beliefs when they log on.

3.4 Changing Beliefs

Influence is measured as a continuous variable from 0 (does not believe disinformation) to 1 (dedicated disinformation believer). To calculate beliefs, we start by assigning a value to each tweet. This value is calculated as follows:

$$Tweet_{value} = type \times similarity_{ij} \times influence_i$$

where *type* $\in \{-1,0,1\}$ indicating *anti-disinformation* (-1), *noise* (0), or *disinformation* (1). i is the user sending the tweets, and $j \in \{\text{followers of i}\}$. Belief is then computed with

$$belief_t = belief_{t-1} + (mean(tweets_{read}) + global_{perc})) \times (1 - belief_{t-1}).$$

The final term is designed to constrain the value between 0 and 1. This also causes diminishing returns in belief (i.e. a dedicated disinformation believer will not become an even greater dedicated believer).

3.5 Agent Based Model Algorithm

The basic time step of our agent based model is presented in Algorithm 1.

```
initialization;
for each time step do
    if start of new week then
        Update similarity matrix;
        Update influence vector (in-degree);
        Update global perception;
    end
    for each user do
        if If user checks Twitter in this time step then
            Get new wake time ;
            Read Tweets ;
            Adjust belief value;
            Create tweets;
            Add retweets to tweets ;
            Create mentions;
            Send tweets/mentions to followers;
            With a given probability create new link with similar user;
            With a given probability create new link with mention author;
        end
    end
end
```

Algorithm 1: Pseudo-code for Twitter disinformation agent based model.

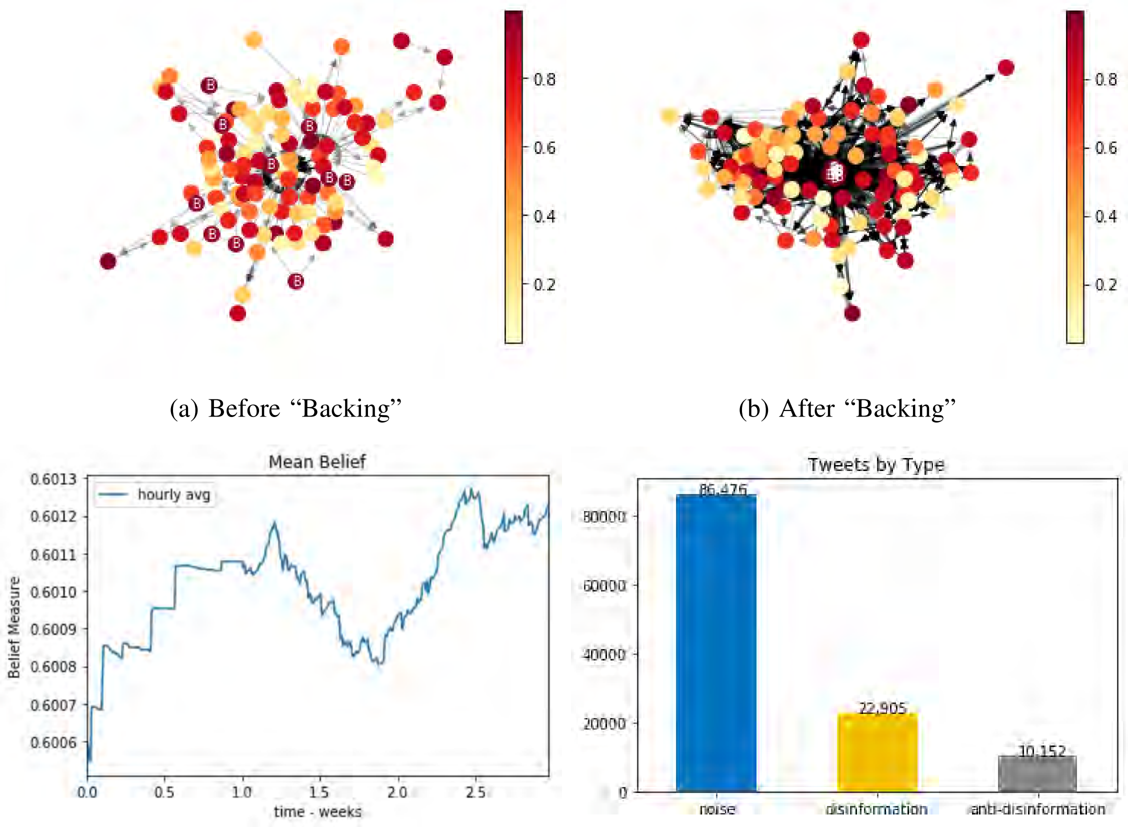
The above algorithm is only slightly modified if the user is a bot/troll or stifler. The stifler will send counter-disinformation *replies* instead of tweets. Bots send disinformation mixed with 20% noise, and also attempt to add links during every session (normal users add link with probability 0.05).

4 EXPLORING KNOWN DISINFORMATION MANEUVERS

In this section we will use *twitter_sim* to explore emergent behavior when bots/trolls conduct known disinformation forms of maneuver. In our study we decided to model *backing* and *bridging*. These two methods were selected from among the 16 maneuvers presented in the BEND framework discussed above.

4.1 Exploring Emergent Behavior when Bots *Back* Influencers

Backing involves following and retweeting influencers to amplify their message. *Back* is defined as "actions that increase the importance of the opinion leader" (Beskow and Carley 2019). These actions can be as simple as following and retweeting the opinion leader. In our experimental design we compare the difference between bots randomly retweeting and following versus targeted backing of the influencers (agents with high in-degree).



(a) Before "Backing"

(b) After "Backing"

(c) Mean Beliefs

(d) Types of Tweets Sent

Figure 1: 100 node scale free network before and after 2.5 weeks of bot *backing* operation and other normal Twitter behavior and network evolution (bots are labeled with "B").

Figure 1 visualizes network topology and belief density before and after 500 time steps ($\sim$ 2.5 weeks). Here we see the natural clustering that tends to occur due to homophily, as well as a limited belief uniformity within clusters. In Figure 1c we see the mean belief fluctuates as bots promote disinformation and *stiflers* suppress the disinformation campaign. Tweets by type are provided in Figure 1d.

We see that bots initially attract other bots to follow them, and slowly get regular users to follow them. At the end of 500 time steps, most bots had 1 or at most 2 normal users following them (as well as 5 to 10 other bots). At face value this mimics what we observe and expect in our empirical bot research.

Given that, at face value, this small network mirrors the behavior we expect from bots that *back* and promote, we ran an experiment on larger networks. We conducted a total of 96 runs of the experiment on 1000 node small world networks with bot percentages that range from 5% up to 20%. The experimental design is is provided in Table 2.

The results are presented in Figure 2. We see that *stiflers* are able to maintain a downward trend on belief in disinformation until the bot percentage exceeds 12%. At a bot percentage of 12% the battle for belief is generally at a stalemate, and at bot percentages greater than 12% the bot campaign begins to build increasing belief in disinformation. We found that random following appears to perform better than targeted following in most scenarios, but never with statistically significant results. The bots that *back* influencers were able to capitalize on the prestige of the influencers, but weren't able to embed in networks

Table 2: Experimental design for backing and bridging.

	Bot Percentage					
	5%	8%	10%	12%	15%	20%
Baseline (random following)	12	12	12	12	12	12
Back (targeted following of influencers)	12	12	12	12	12	12
Bridge with Random Following	12	12	12	12	12	12
Bridge with Targeted Following	12	12	12	12	12	12

and gain followers as fast as the random following scenario. Because the bots that randomly follow were able to embed in local networks and gain followers faster, they achieved at least parity with the bots that *back* influencers.

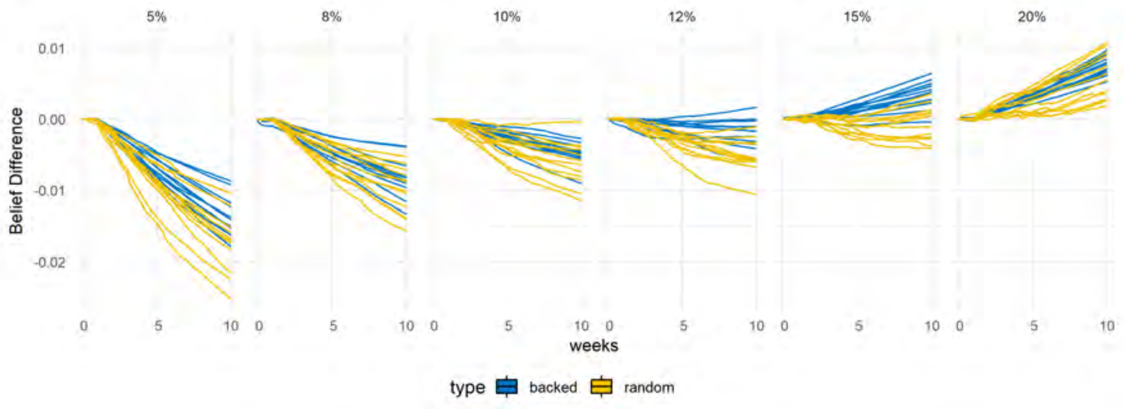


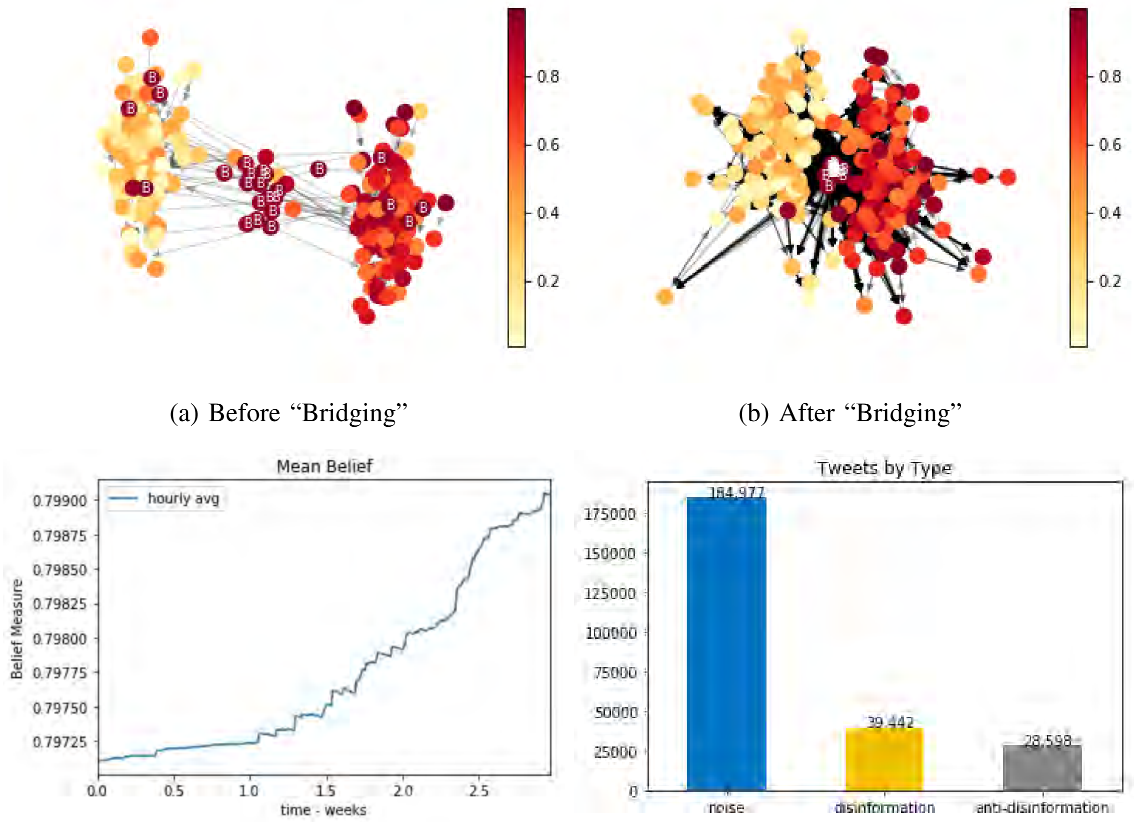
Figure 2: Results of backing on 96 runs with random and targeted *backing*.

4.2 Exploring Emergent Behavior when Bots ”Bridge” Communities

Bridging involves building links between two communities in order to introduce the ideas of one into the other. Bridge is defined as “actions that build a connection between two or more groups” (Beskow and Carley 2019). Our team has observed this behavior in political conversations in the United States and in Europe. The perpetrators identify a target community that they want to influence with the ideas and norms of a separate community. For example, in US political events our team has observed efforts to connect far-right groups with religious communities. The perpetrators first attempt to embed into the target community. Once embedded, they introduce ideas and create connections between the two groups.

It is important to note that the *backing* algorithm was not changed for *bridging*. The only difference is the input network topology consisted of a two communities instead of one. In the process we found that if bots conduct either random or targeted *backing* when given two communities, they will inevitably “bridge” those communities. This is an important confirmation for us. Even when researchers observe bots with high betweenness and make assumptions of an underlying intention to bridge, it may just be that a bot that is backing influencers has been intentionally or unintentionally oriented on two or more communities.

Figure 3 visualizes network topology and belief density before and after 500 time steps ($\sim$ 2.5 weeks). Here we see two separate groups, one that strongly believes in the disinformation message, and one that does not. The bots have been inserted in the middle with a single following tie to each of the two communities. In Figure 3b we observe the network after 500 time steps. We see that even after only $\sim$ 2.5 weeks, the bots have already started to bring the two communities together and are starting to introduce the ideas of the one group into the other group. In Figure 3c we see that the mean beliefs of the target community



(a) Before "Bridging"

(b) After "Bridging"

(c) Mean Beliefs of Target Community

(d) Types of Tweets Sent

Figure 3: 100 Node scale free network before and after 500 hours of a bot *bridging* operation and other normal Twitter behavior and network evolution (bots are labeled with "B").

are already starting to increase, demonstrating that they are beginning to believe in the disinformation. In Figure 3d we see the types of tweets sent during the ~ 2.5 weeks.

We see that *bridging* does bring the two communities together, and the *siflers* that are present in the target community are not able to prevent increasing belief in the disinformation message. We also see the network topology evolve, bringing the two groups together with bots having high betweenness centrality.

Given that, at face value, this small network mirrors the behavior we expect from bots that *bridge* two communities, we set to run an experiment on larger networks. We conducted a total of 96 runs of the experiment on 2000 node small world networks with bot percentages that range from 5% up to 20%. The target community consists of a 1000 node scale free network with initial beliefs distributed between 0 and 0.5, and the "host" network consists of a 1000 node scale free network with initial beliefs distributed between 0.5 and 1. The experimental design is provided in Table 2.

The results are presented in Figure 4. The results show the average belief of the target community (not the whole network). At a bot percentage of 12% the battle for belief is generally at a stalemate, and at bot percentages greater than 12% the bot campaign begins to build increasing belief in disinformation. We see that random and targeted following appear to be the same. This makes sense given the observation made above, and validates that a bot that is programmed to either back influencers or randomly promote, when intentionally or unintentionally pointed at two communities, will automatically begin *bridging* those communities and communicating beliefs and norms between them.

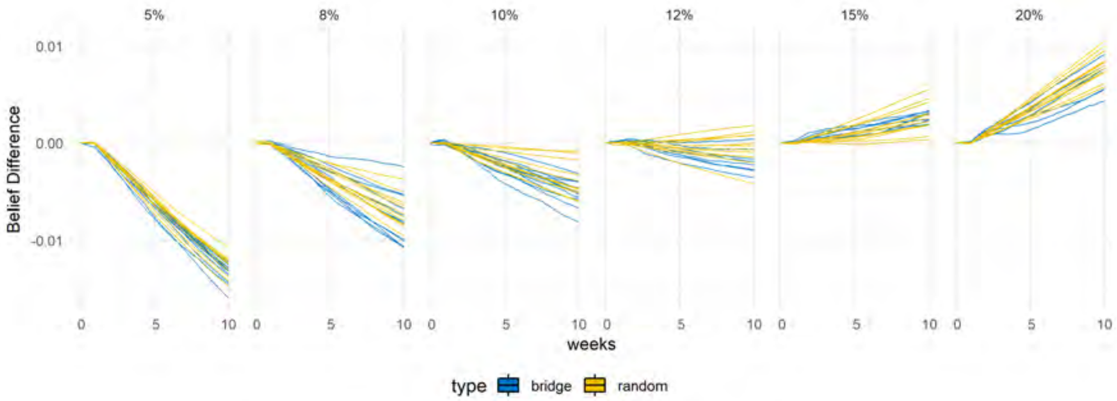


Figure 4: Results of *bridging* on 96 runs with random and targeted *backing*.

5 VALIDATION

Most rumor propagation models measure the spread of the rumor, not the outcome in people's beliefs. This is primarily because it is easy to measure and validate the spread of information, but is nearly impossible to measure and validate the beliefs of an online community. While proxies may exist in some cases (for example stance can be inferred by hashtags such as #ClimateHoax or #GunControlNow), these proxies may be weakly correlated and are subject to manipulation (i.e. hashtag latching). While not able to validate belief, we still model belief based on exposure as our outcome variable, paired with a simulation that closely replicated the Twitter environment with strong face validity.

Given these limitations, we focused our validation efforts on ensuring we accurately modeled the behavior that we empirically see in Twitter. The focus of this validation was on estimating the distribution of the inter-arrival time of a Twitter Users' *session* (single log-on episode) and the activity of a single *session*. We also sought to make sure our distribution of original tweets, retweets, and replies replicated the distribution seen in typical Twitter streams.

To validate these metrics we collected three separate Twitter Streams. We collected tweets associated with followers of all US Congressional politicians and congressional candidates of the 2018 mid-term elections. From this we sampled 10M tweets from 56,908 users. The second data set was from tweets associated with users discussing the 2018 Swedish mid-term elections (18M Tweets from 101,260 users). The final data set was a sample of tweets from the 1% Twitter Sample (11M Tweets from 20,144 users). While there is evidence that this 1% sample is not random (Morstatter et al. 2013), we still felt that this would give a data perspective that was not tied to a political event/process.

For each data set we collected the last 200-600 tweets of the users. We then tried to segment the tweet timeline into *sessions*. For the sake of this paper we determined that if any tweet inter-arrival time was greater than 15 minutes, that it constituted a new *session*. If tweet inter-arrival time was less than 15 minutes, we considered it as the same *session*. Having segmented the user timelines, we were then able to calculate tweet inter-arrival time, session inter-arrival time, the user rate (sessions per day), and the tweets per session. Descriptive statistics are provided in Table 3.

We do want to highlight that numerous bots are present in this data. Using the bot-hunter tool (Beskow and Carley 2018) with a threshold of 0.6, the US midterm election data contains 38.9% bots, the Swedish Election data contains 45.7% bots, and the 1% Sample contains 12.8% bots. Some of these bots tweet almost constantly, and therefore all tweets in our sample of their timelines were considered a single *session*. These accounts also can tweet at the speed of algorithms, which is why the minimum inter-arrival time for all three data sets rounds to 0.

Given at least one mention, the mean mentions for *Sample* and *Sweden* were approximately 1.5 mentions, while US elections was 3 mentions. Note that for all three data sources, the 75 percentile was still 1

Table 3: Empirical validation of Twitter user 'sessions'.

		Mean	Min	1st QT	Median	3rd QT	Max
Twitter 1% Sample	Tweet Interarrival (hrs)	7.15	0	0.011	0.08	1.39	6.82 yrs
	Session Interarrival (hrs)	16.5	0.167	0.68	2.1	10.6	6.82 yrs
	User Rate (Sessions/Day)	9.73	0.00027	0.0086	0.032	0.15	111K
	Tweets per Session	2.3	1	1	1	2	> 3200
Swedish Election Data	Tweet Interarrival (hrs)	33.5	0	0.027	0.3	10.3	10 yrs
	Session Interarrival (hrs)	60	0.167	1.14	7.31	27.3	10 yrs
	User Rate (Sessions/Day)	41.2	0.00024	0.004	0.02	0.166	1036K
	Tweets per Session	1.76	1	1	1	3	> 3200
US 2018 Midterm Election Data	Tweet Interarrival (hrs)	61.9	0	0.041	1.27	22	9.14 yrs
	Session Interarrival (hrs)	97.8	0.167	1.82	13.3	46.6	9.14 yrs
	User Rate (Sessions/Day)	3.7	0.00024	0.0018	0.0071	0.0447	20K
	Tweets per Session	1.57	1	1	1	1	> 3200

mention, indicating a highly skewed distribution. We found that 40-50% of tweets are tweets and 20-25% are a replies. We also found that approximately 70-80% contain a mention (note that all retweets contain a mention to the originator).

6 CONCLUSION

In this paper we've presented the *twitter_sim* ABM designed for exploring the explicit actions users make in Twitter and capturing the varied actions of malicious agents like bots/tolls. We've demonstrated the use of this model in exploring the emerging behavior of specific disinformation maneuvers. Finally, we've validated some of the key variables in the model from empirical Twitter data.

Using *twitter_sim* to explore *backing* and *bridging* campaigns demonstrates that bots are not effective if they don't embed in networks and gain a following, even if they are amplifying the messages of influencers. In the *bridging* experiment we observe that *bridging* can occur simply by pointing bots that are programmed to *back* at multiple communities. In the process of *backing* key individuals, they will *bridge* the network.

twitter_sim is an extremely adaptable and scalable agent based model that fills some key voids for those studying disinformation, information diffusion, or even marketing more generally. It also offers promise for those seeking a model to create a virtual social media environment for training environments.

ACKNOWLEDGMENT

This work was supported in part by the Office of Naval Research (ONR) Multidisciplinary University Research Initiative Award N000140811186 and Award N000141812108, and the Center for Computational Analysis of Social and Organization Systems (CASOS). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the ONR or the U.S. government.

REFERENCES

Beskow, D., and K. M. Carley. 2018. "Introducing Bothunter: A Tiered Approach to Detection and Characterizing Automated Activity on Twitter". In *Proceedings of the International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, edited by R. Thompson, C. Dancy, A. Hyder, and H. Bisgin, Volume 10899 of *Lecture Notes in Computer Science*. Cham, Switzerland: Springer.

Beskow, D. M., and K. M. Carley. 2019. "Social Cybersecurity: An Emerging National Security Requirement". *Military Review* 99(2):117–127.

Bessi, A., and E. Ferrara. 2016. “Social Bots Distort the 2016 US Presidential Election Online Discussion”. *First Monday* 21(11-7).

Carley, K. M. 1990. “Group Stability: A Socio-cognitive Approach”. *Advances in Group Processes* 7(1):44.

Carley, K. M., M. K. Martin, and B. R. Hirshman. 2009. “The Etiology of Social Change”. *Topics in Cognitive Science* 1(4):621–650.

Chavoshi, N., H. Hamooni, and A. Mueen. 2016. “DeBot: Twitter Bot Detection via Warped Correlation.”. In *Proceedings of the International Conference on Data Mining*, edited by R. Stahlbock and G. M. Weiss, 817–822. Red Hook, New York: Computer Science Research, Education and Applications Press.

Daley, D. J., and D. G. Kendall. 1965. “Stochastic Rumours”. *Institute of Mathematics and Its Applications Journal of Applied Mathematics* 1(1):42–55.

Ferrara, E., O. Varol, C. Davis, F. Menczer, and A. Flammini. 2016. “The Rise of Social Bots”. *Communications of the Association for Computing Machinery* 59(7):96–104.

Friedkin, N. E. 2006. *A Structural Theory of Social Influence*, Volume 13. Cambridge University Press.

Hagberg, A., P. Swart, and D. S Chult. 2008. “Exploring Network Structure, Dynamics, and Function Using Networkx”. Technical report, Los Alamos National Lab (LANL), Los Alamos, NM (United States).

Maki, D. P., and M. Thompson. 1973. “Mathematical Models and Applications: With Emphasis on the Social Life, and Management Sciences”. Technical report.

McPherson, M., L. Smith-Lovin, and J. M. Cook. 2001. “Birds of a Feather: Homophily in Social Networks”. *Annual Review of Sociology* 27(1):415–444.

Morstatter, F., J. Pfeffer, H. Liu, and K. M. Carley. 2013. “Is the Sample Good Enough? Comparing Data from Twitter’s Streaming API with Twitter’s Firehose”. In *Proceedings of the Seventh International Association for the Advancement of Artificial Intelligence Conference on Weblogs and Social Media*, edited by E. Kiciman, N. B. Ellison, B. Hogan, P. Resnick, and I. Soboroff. Palo Alto, California: Association for the Advancement of Artificial Intelligence Press.

Serrano, E., C. A. Iglesias, and M. Garijo. 2015. “A Novel Agent-Based Rumor Spreading Model in Twitter”. In *Proceedings of the 24th International Conference on World Wide Web*, WWW ’15 Companion, 811–814. New York, NY, USA: Association for Computing Machinery.

Tripathy, R. M., A. Bagchi, and S. Mehta. 2010. “A Study of Rumor Control Strategies on Social Networks”. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, Conference of Information and Knowledge Mangement ’10, 1817–1820. New York, NY, USA: Association for Computing Machinery.

Wang, C., Z. X. Tan, Y. Ye, L. Wang, K. H. Cheong, and N.-g. Xie. 2017. “A Rumor Spreading Model Based on Information Entropy”. *Scientific reports* 7(1):9615.

Weng, L., A. Flammini, A. Vespignani, and F. Menczer. 2012. “Competition Among Memes in a World with Limited Attention”. *Scientific Reports* 2:335.

Zanette, D. H. 2002. “Dynamics of Rumor Propagation on Small-world Networks”. *Physical Review E* 65(4).

AUTHOR BIOGRAPHIES

DAVID M. BESKOW is a PhD candidate in the School of Computer Science at Carnegie Mellon University. He holds a BS from the United States Military Academy in civil engineering and an MS from the Naval Postgraduate School in operations research. As an operations research and systems analyst (ORSA), Beskow served as an assistant professor at West Point and as an ORSA analyst at the U.S. Army. Beskows current research develops machine learning algorithms to detect and characterize online disinformation. His email address is dnbeskow@gmail.com.

KATHLEEN M. CARLEY is a professor of societal computing in the School of Computer Science at Carnegie Mellon University, an IEEE Fellow, the director of the Center for Computational Analysis of Social and Organizational Systems (CASOS), and the CEO of Netanomics. She is the 2011 winner of the Simmel Award from the International Network for Social Network Analysis and the 2018 winner of the National Geospatial-Intelligence Agency Academic Award from GEOINT. Her email address is kathleen.carley@cs.cmu.edu.