Modeling Malicious Behaviors and Fake News Dissemination on Social Networks

Kento Yoshikawa, Masatsugu Ichino, Hiroshi Yoshiura

Modeling Malicious Behaviors and Fake News Dissemination on Social Networks

Kento Yoshikawa, Masatsugu Ichino, and Hiroshi Yoshiura

The University of Electro-Communications, Tokyo, Japan
{k-yoshikawa,yoshiura}@uec.ac.jp, ichino@inf.uec.ac.jp

Abstract. As social media has become widely used, fake news has become a serious problem. A representative countermeasure is fake news detection. However, this countermeasure is not sufficient because people using social media tend to ignore facts that contradict their beliefs. To develop effective countermeasures, it is necessary to clarify the influence of fake news and the nature of its dissemination from the perspective of communication. In this paper, we propose two models explaining the dissemination of opinions about fake news: one in which the presence of the ground truth is assumed and one in which it is not assumed. In both models, an attacker disseminates fake news by imitating or hijacking target accounts. In evaluations on real-world social networks, the model in which the ground truth is assumed demonstrates that, contrary to our expectations, account imitation is a more harmful attack than account hijacking. The model in which ground truth is not assumed demonstrates that both account imitation and account hijacking are harmful attacks.

Keywords: Fake news · Social network · Opinion dissemination · Account imitation · Account hijacking

1 Introduction

The platforms on which people receive and disseminate information are changing from mass media (e.g., newspapers, television) to social media (e.g., Facebook, Twitter). Social media enable people to get information easily and to disseminate information rapidly. As a result, misleading news, including deceptive news, has become widespread. This misleading news, i.e., fake news, negatively affects individuals and society [11].

A representative countermeasure against fake news is fake news detection using texts and images [14], the speed at which the news spreads [13], and the reliability of people reporting the news [15]. However, widespread communication on social media has amplified the echo chamber effect in which one's beliefs are strengthened through interactions with like-minded individuals [8]. It has also amplified the backfire effect in which facts that contradict one's beliefs are rejected [8]. Fake news detection is thus insufficient.

Developing effective countermeasures requires clarifying the influences of fake news and the nature of its dissemination from the perspective of communication.

Previous models related to this aim describe opinion dissemination among people using social media (hereinafter "users") [3,4,7,10,12]. However, these models are insufficient because they do not take malicious attacks into account.

In this paper, we present two models explaining the dissemination of opinions about fake news generated by malicious users (hereinafter "attackers") on social networks. We used them to clarify the influence and properties of fake news. Our contributions to countermeasures against fake news are summarized as follows.

- We present two opinion dissemination models that users disseminate news in the presence of attackers. In the AAT-Based Model, attackers and users disseminate news when the ground truth is assumed while in the Trust-Based Model, they disseminate news when the ground truth is not assumed. In both models, attackers intentionally facilitate the spread of fake news by imitating or hijacking the accounts of target users.
- Using the AAT-Based Model on real-world social networks, we reveal an unexpected result that account imitation is a more harmful attack than account hijacking. For example, with account imitation, the attack is always effective, and 1% of attackers deter more than 80% of the users from making up their opinions. With account hijacking, the attack sometimes fails.
- Using the Trust-Based Model on real-world social networks, we reveal that account imitation and account hijacking can cause two outcomes even though there are only 1% of attackers: attackers can (1) facilitate the spread of opinions that support them and (2) suppress the spread of targets' opinions.

2 Related Work

2.1 Modeling Opinion Dissemination

Previous work proposes models explaining how users in social networks disseminate opinions. These models can be classified into two types: those that assume the presence of the ground truth (i.e., the fact that supports or contradicts an opinion) [4, 7] and those that do not [3, 10, 12].

The former type describes the dissemination of the correct (or incorrect) opinion that matches (or mismatches) the ground truth. Glinton et al. modeled the "opinion sharing problem" in which users share the correct opinion [4]. Pryymak et al. improved the precision of Glinton's model by developing an opinion dissemination model, AAT (Autonomous Adaptive Tuning) [7].

The latter type does not consider whether an opinion matches the ground truth and simply describe how people change their opinions due to communication. DeGroot proposed a model in which users update their opinions based on the fixed weighted average of the importance of their friends [3]. As extensions of DeGroot's model, Tsang and Larson modeled an opinion transition that diverse opinions converge to a few major opinions by adding people who never change their opinions [12]. Sasahara et al. modeled the echo chamber effect by formulating the disconnection with those who have different opinions [10].

In this paper, we adopt Pryymak's model [7] as the model in which the ground truth is assumed because it can accurately share correct opinions. We also adopt Tsang and Larson's model [12] as the model in which the ground truth is not assumed because it can better reflect opinion formation by updating not only opinions but also the importance of friends while other models of this type cannot. We incorporate malicious attacks (described in Section 2.2) into these two models (Section 3).

2.2 Malicious Attacks Facilitating Spread of Fake News

We first define "fake news" in this paper as deceptive news that attackers not only generate but also actively spread by deceiving users [11]. Malicious attacks on social networks that correspond to this definition are account imitation and account hijacking [5, 9]. In account imitation, the attacker generates accounts similar to the target accounts in terms of account names, photos, and texts [5,9]. In account hijacking, the attacker hijacks the target accounts through phone calls, email, and linking social media with external applications [5, 9]. Although such attacks are rampant on social media, there has been no work to evaluate the effects of spreading fake news at present.

3 Modeling Malicious Behaviors and Opinion Dissemination for Fake News

3.1 Overview

The goal of an attacker is to convince users that the information benefiting the attackers is correct (e.g., political propaganda). To achieve this goal, the attacker disseminates "misinformation" and then performs account imitation or hijacking as described in Section 2.2 in order to facilitate its spread (① in Fig.1). We analyze the dissemination of misinformation using Pryymak's model [7] (hereinafter "AAT-Based Model") and Tsang and Larson's model [12] (hereinafter "Trust-Based Model")(② in Fig.1). The AAT-Based Model assumes the presence of the ground truth while the Trust-Based Model does not. In the AAT-Based Model, "misinformation" means opinions that misidentify real news as fake news or misidentify fake news as real news. In the Trust-Based Model, "misinformation" means opinions about the news that the attacker shares with users: fake positive opinions that are opposite to the targets' negative opinions or fake negative opinions that are opposite to the targets' positive opinions.

3.2 Problem Formulation

We consider graph $G(U, E)$ to be a social network, where $U = \{u_1, \ldots, u_N\}$ is the set of users (i.e., nodes), and E is the set of friendships among all users (i.e., links). Each user u_i $(i = 1, \ldots, N)$ has M_i $(1 \leq M_i \leq N - 1)$ friends in $F_i = \{u_{i_1}, \ldots, u_{i_j}, \ldots, u_{i_{M_i}}\} \subset U$, where F_i is the set of user u_i's friends and

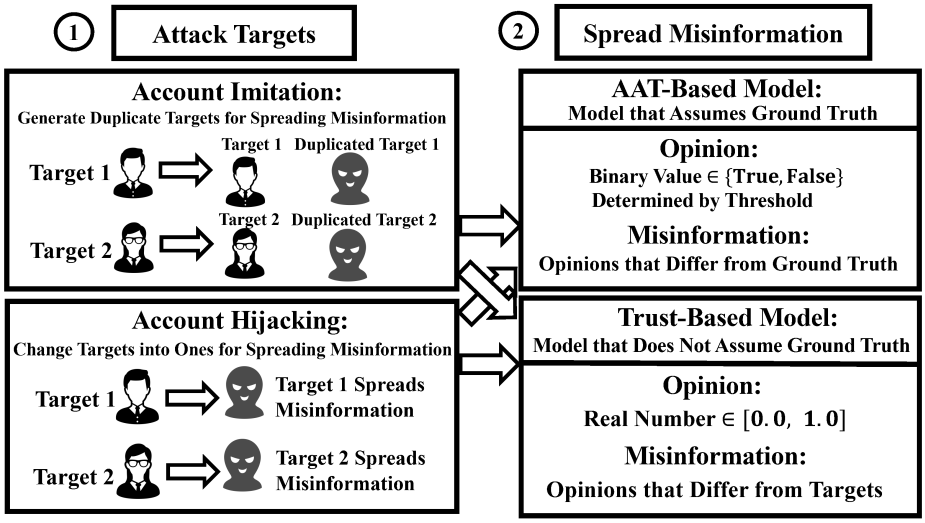

Fig. 1. Overall illustration of proposed models

u_{i_j} is the jth friend of user u_i. In accordance with the models of Glinton et al. [4] and Pryymak et al. [7], each user communicates their opinions with the set of their friends. Each user u_i has an opinion $o_i \in [0.0, 1.0]$ and an importance vector of friends $\boldsymbol{W}_i = (w_{i_1}, \ldots, w_{i_j}, \ldots, w_{i_{M_i}})$. Each user u_i updates his or her current opinion o_i^k to o_i^{k+1} by applying

$$o_i^{k+1} = f(o_i^k, \boldsymbol{W}_i^k, \boldsymbol{O}_i^k), \tag{1}$$

where k is the current opinion update step, $\boldsymbol{W}_i^k = (w_{i_1}^k, \ldots, w_{i_j}^k, \ldots, w_{i_{M_i}}^k)$ is the current importance vector of friends, and $\boldsymbol{O}_i^k = (o_{i_1}^k, \ldots, o_{i_j}^k, \ldots, o_{i_{M_i}}^k)$ is the current opinion vector of user u_i's friends u_{i_j} $(j = 1, \ldots, M_i)$. Function $f(o_i^k, \boldsymbol{W}_i^k, \boldsymbol{O}_i^k)$ is embodied in Section 3.4. The attacker performs account imitation or account hijacking on the set of targets $B \subset U$ embodied in Section 3.3.

- **Account Imitation:** The attacker generates duplicate target accounts and then adds them to a social network. The duplicated accounts spread misinformation among the friends of the targets. This attack represents the situation in which those who spread misinformation increase while those who spread correct information remain unchanged.
- **Account Hijacking:** The attacker succeeds in hijacking target accounts with some probability. The hijacked accounts spread misinformation among the friends of the targets. The probability of success varies from target to target depending on their security against hijacking. This attack represents the situation in which those who spread misinformation increase while those who spread correct information decrease by the same number.

By applying these attacks in the AAT-Based Model and Trust-Based Model, we clarify the influence of attacks on opinion dissemination in three cases: no malicious attacks, account imitation, and account hijacking.

3.3 Attacker Behaviors

We consider a situation in which an attacker performs account imitation or account hijacking against $|B|$ influencers (i.e., people who have many friends or are reliable) as targets. We assume that the attacker knows the number of friends for each target [1]. We next embody account imitation, account hijacking, and misinformation in the AAT-Based Model and Trust-Based Model.

- **AAT-Based Model**
  We define the ground truth as $z \in \{\text{True}, \text{False}\}$. If $z = \text{True}$, the news is real news (e.g., the real news "Wildfires broke out in California in 2020", hereinafter "News California"). If $z = \text{False}$, the news is fake news (e.g., the fake news "5G can cause the COVID-19 infection"). We consider the targets to be users whose number of friends is within the top $m\%$ among all users in a social network $G(U, E)$ (i.e., influencers). We define misinformation as $\bar{z}$, which is opposite to the ground truth z.
  - **Account Imitation:** In a random opinion update step k_{rand}, the attacker generates duplicate target accounts b' ($b' = 1, \ldots, |B|$), which have the reliability t_{i_b} and doubt f_{i_b}. Reliability t_{i_b} is the degree to which each user u_i believes targets b. Doubt f_{i_b} is the degree to which each user u_i doubts targets b. The duplicated accounts b' share opinions $o_{b'} = \bar{z}$ with the targets' friends F_b.
  - **Account Hijacking:** The attacker succeeds in hijacking target accounts b with probability p_b. The hijacked accounts b share opinions $o_b = \bar{z}$ with F_b.
- **Trust-Based Model**
  We consider targets to be users whose number of friends is within the top $m\%$ among extremists, who have extreme opinions (i.e., extremists among influencers). We define extreme opinions as opinion $o = 0$ or 1 (e.g., radical conservatives or radical liberals). Following Tsang and Larson [12], extremists never change their opinions. We define misinformation as $\bar{o}_b = 1 - o_b$, which is opposite to the targets b's opinions $o_b \in \{0, 1\}$.
  - **Account Imitation:** The attacker generates duplicate target accounts b' in step k_{rand}. The duplicated accounts b' share $o_{b'} = \bar{o}_b$ with F_b.
  - **Account Hijacking:** The attacker succeeds in hijacking target accounts b with p_b. The hijacked accounts b share $o_b = \bar{o}_b$ with F_b.

3.4 Opinion Formulation

We embody the opinion update (i.e., the function $f(o_i^k, \boldsymbol{W}_i^k, \boldsymbol{O}_i^k)$) in the AAT-Based Model and Trust-Based Model.

- **AAT-Based Model**
  The opinion of each user u_i is the subjective probability that he or she

[1] This assumption is realistic because we can know the number of friends by accessing user profile pages on social media (e.g., the number of followers on Twitter).

believes the news is real. We denote the opinion as $o_i = P_i(z = \text{True})$. We also denote $P_i(z = \text{False}) = 1 - P_i(z = \text{True})$ as the subjective probability that each user u_i believes the news is fake. If opinion $P_i(z = \text{True}) \geq$ a threshold σ $(0.5 < \sigma < 1.0)$, each user u_i shares opinion $o_i = \text{True}$ with each friend u_{i_j} $(j = 1, \ldots, |F_i|)$, which means that each user u_i tells their friends the news is real (e.g., each user tells their friends "News California" is real). If opinion $P_i(z = \text{True}) \leq 1-\sigma$, each user u_i shares opinion $o_i = \text{False}$ with each friend u_{i_j}, which means that each user u_i tells the news is fake (e.g., each user tells "News California" is fake although it is real). If $1 - \sigma < P_i(z = \text{True}) < \sigma$, users do not share their opinions. The higher the threshold σ, the more careful users share their opinions; the lower the threshold, the more willing users share their opinions. If each user u_i receives opinions $o_{i_j} \in \{\text{True}, \text{False}\}$ from their friends u_{i_j}, he or she updates current opinion $P_i^k(z)$ to $P_i^{k+1}(z|o_{i_j})$ using (2), which is based on Bayes' theorem:

$$P_i^{k+1}(z|o_{i_j}) = f(o_i^k, \boldsymbol{W}_i^k, \boldsymbol{O}_i^k) = \frac{P_i(o_{i_j}|z)P_i^k(z)}{\sum\limits_{z \in \{\text{True},\text{False}\}} P_i(o_{i_j}|z)P_i^k(z)}, \tag{2}$$

where $\boldsymbol{W}_i^k = (t_{i_1}, \ldots, t_{i_j}, \ldots, t_{i_{M_i}}; f_{i_1}, \ldots, f_{i_j}, \ldots, f_{i_{M_i}})$, $t_{i_j} = P_i(o_{i_j} = \text{True}|z = \text{True}) \in [0.0, 1.0]$ is reliability, and $f_{i_j} = P_i(o_{i_j} = \text{True}|z = \text{False}) \in [0.0, 1.0]$ is doubt, as described in Section 3.3. In the AAT [7], $t_{i_j} = t_{i_{j'}}$ and $f_{i_j} = f_{i_{j'}}$, where $\forall(j, j') \in F_i$ (i.e., a user believes or doubts friends equally). Pryymak et al. pointed out that t_{i_j} and f_{i_j} should differ for each friend (i.e., a user believes or doubts friends differently) [7]. To distinguish whether a user is an influencer (i.e., those who are reliable) or not, we followed Pryymak's remark. The opinion update process continues until the current opinion update step k exceeds the maximum opinion update step K.

- **Trust-Based Model**

  If each user u_i receives opinion $o_{i_j} \in [0.0, 1.0]$ from friends u_{i_j}, he or she updates current opinion $o_i^k \in [0.0, 1.0]$ to o_i^{k+1} using (3), which is based on the weighted average:

$$o_i^{k+1} = f(o_i^k, \boldsymbol{W}_i^k, \boldsymbol{O}_i^k) = \frac{w_{ii}o_i^k + \sum_{j \in F_i} w_{i_j}^k o_{i_j}^k}{w_{ii} + \sum_{j \in F_i} w_{i_j}^k}, \tag{3}$$

  where $\boldsymbol{W}_i^k = (w_{i_1}^k, \ldots, w_{i_j}^k, \ldots, w_{i_{M_i}}^k)$ is the current importance vector of friends, $\boldsymbol{O}_i^k = (o_{i_1}^k, \ldots, o_{i_j}^k, \ldots, o_{i_{M_i}}^k)$ is the current opinion vector of friends, and w_{ii} is the importance of the user himself or herself. Each user u_i also updates the elements of the current importance vector of friends using (4):

$$w_{i_j}^{k+1} = \frac{w_{i_j}^k + rT(o_i^k, o_{i_j}^k)}{1 + r}, \tag{4}$$

  where r is the learning rate. The higher it is, the less likely users accept opinions that differ from their opinions. $T(o_i^k, o_{i_j}^k)$ is the reliability function:

$$T(o_i^k, o_{i_j}^k) = \exp\left(-\frac{(o_i^k - o_{i_j}^k)^2}{h}\right), \tag{5}$$

which represents that users are more likely to rely on someone who has similar values and less likely to rely on someone who has different values. The h is a value representing empathy for opinions. The higher it is, the more likely users accept friends' opinions. Each user u_i updates their opinions using (3) and then updates importance using (4) [12].
Tsang and Larson proposed three methods for initializing friend's importance: the same value for all friends, a normal distribution, and the number of friends [12]. Since the targets are influencers (i.e., users have many friends), we use the third initialization method and initialize importance using

$$w_{i_j} = \frac{d_{i_j}}{d_i} \qquad w_{ii} = \frac{d_i}{d_i} = 1, \tag{6}$$

where d_i is the number of user u_i's friends and d_{i_j} is the number of friends each friend u_{i_j} has. All users can always share their opinions. Opinion updating continues until the change in the opinions of all users becomes less than a small threshold ε or the current opinion update step k exceeds the maximum opinion update step K.

4 Algorithms for Malicious Behaviors and Opinion Dissemination Models for Fake News

4.1 Algorithms for Opinion Updating

In the AAT-Based Model, each user updates his or her opinion and shares it in accordance with threshold σ and $1 - \sigma$ (Algorithm 1). In the Trust-Based Model, each user updates his or her opinion before updating the importance of their friends (Algorithm 2). If attacks terminate or do not exist from the beginning, Algorithm 3 in Section 4.2 is not executed.

Algorithm 1 Opinion Updating in AAT-Based Model

Initialize social network $G(U, E)$, ground truth z, threshold σ, opinion, reliability and doubt.
while $k \leq K$ **do**
  for $i = 1$ **to** $|U|$ **do**
    if an attack exists, **then**
      Attacker executes account imitation or account hijacking in accordance with Algorithm 3.
    if current opinion $o_i^k = P_i^k(z = \text{True}) \geq \sigma$, **then**
      Send $o_i^k = \text{True}$ to all friends belonging to the set F_i
    else if $o_i^k = P_i^k(z = \text{True}) \leq 1 - \sigma$, **then**
      Share $o_i^k = \text{False}$ with all friends in F_i.
    Update current opinion $o_i^k = P_i^k(z = \text{True})$ to $o_i^{k+1} = P_i^{k+1}(z = \text{True})$ using (2).
  Set $k = k + 1$

Algorithm 2 Opinion Updating in Trust-Based Model

Initialize social network $G(U, E)$, opinions, importance, learning rate r, empathy h, and threshold ε.
while $k \leq K$ **or** Differences in opinions of all users $\leq \varepsilon$ **do**
  for $i = 1$ **to** $|U|$ **do**
    if an attack exists, **then**
      Attacker executes account imitation or account hijacking in accordance with Algorithm 3.
    Update current opinion o_i^k to o_i^{k+1} using (3).
    Update current importance w_{ij}^k for all friends belonging to the set F_i using (4) and (5).
  Set $k = k + 1$

4.2 Algorithm for Attacker Bahaviors

As shown in Algorithm 3, an attacker executes account imitation or account hijacking on targets. In the AAT-Based Model, misinformation is $\bar{z}$, which differs from ground truth $z \in \{\text{True}, \text{False}\}$. In the Trust-Based Model, misinformation is $\bar{o}_b = 1 - o_b$, which differs from the targets' opinions $o_b \in \{0, 1\}$.

Algorithm 3 Attacker Behaviors

Initialize update step k_{rand} to start attacking and update step T at which attacks are detected.
for target account $b = 1$ **to** $|B|$ **do**
  if Attack = "Account Imitation", **then**
    Generate set of duplicate target accounts B' at opinion update step k_{rand}.
    Duplicated accounts $b'(b' = 1, \ldots, |B|)$ spread misinformation in accordance with AAT-Based Model or Trust-Based Model.
  else if Attack = "Account Hijacking", **then**
    Hijack targets b with probability p_b.
    Hijacked targets b spread misinformation in accordance with AAT-Based Model or Trust-Based Model.
if current opinion update step $k \geq T$, **then**
  Terminate account imitation or account hijacking.

5 Evaluation

5.1 Overview

Through experiments, we clarify the influence of attacks on opinion dissemination by comparing the performance in three cases (i.e., no malicious attacks, account imitation, and account hijacking) for the AAT-Based Model and Trust-Based Model. The performance of the AAT-Based Model is evaluated in terms of accuracy A, inaccuracy I, and undetermined UD, where $A = \frac{\sum_{i=1}^{N} |o_i = z|}{N}$ (the

rate of users having opinion o_i the same as ground truth z), $I = \frac{\sum_{i=1}^{N} |o_i = \bar{z}|}{N}$ (the rate of users having opinion o_i different from ground truth z), and $UD = 1 - A - I$. The performance of the Trust-Based Model is visualized using heatmaps of the opinion distribution among users along with the opinion update steps. The models and experiments were implemented using Python 3.8.5.

5.2 Datasets and Experimental Settings

We used the publicly available Facebook and Twitter social network datasets, which we downloaded from the website "Stanford Network Analysis Project" [6]. These datasets consist of nodes (i.e., users) and links (i.e., friendships among users). Each node has attributes (e.g., age, gender) and connects with other nodes reciprocally or partially reciprocally. The Facebook dataset consists of $4,039$ users and $88,234$ reciprocal links, with each user u_i having $\bar{F}_i = 44$ friends on average. The Twitter dataset consists of $81,306$ users and $1,768,149$ partially reciprocal links, with $\bar{F}_i = 33$ friends. We used only the links in the experiments.

The common settings for both models were as follows. We set the attackers as 1% of the users in the datasets. For account hijacking, the probability p_b that an attacker succeeds in hijacking target accounts b was a uniform distribution with a range of $[0.0, 1.0]$. The attacks were detected at 10 hours or 20 hours after they started in the real world [1,2]. We converted these hours into the opinion update step T. The maximum opinion update step was 3,000, which corresponds to 24 hours. We conducted 50 simulations for each social network (Facebook, Twitter) and attack method (account imitation, account hijacking).

For the AAT-Based Model, the targets were the users whose number of friends was within the top 1%. The initial opinion value was a normal distribution $\mathcal{N}(\text{mean} = 0.5, \text{standard deviation} = 0.15)$, and the initial values of users' reliability and doubt were a uniform distribution with a range of $[0.0, 1.0]$. The threshold σ to share opinions was 0.8. These parameter settings are in accordance with the AAT [7]. When a simulation terminated, the attack was considered to be a failure if $A \geq 80\%$ and to be a success if $I \geq 80\%$.

For the Trust-Based Model, the targets were the extremists (i.e., those who never change their opinions) whose number of friends was in the top 1%. Extremists were assumed to account for 10 % for each extreme opinion $o \in \{0, 1\}$. The remaining 80% of the user opinions were initialized with a uniform distribution with a range of $(0.0, 1.0)$. We set learning rate $r = 1.5$, empathy $h = 0.01$, and termination criterion $\varepsilon = 0.001$. These settings follow Tsang and Larson [12].

5.3 Results and Discussion

- **AAT-Based Model**
  When there were no attacks, accuracy A for real news was higher than that for fake news (i.e., people can more accurately perceive that real news is real than they can perceive that fake news is fake).
  With account imitation, accuracy A decreased substantially, and undetermined UD increased substantially (i.e., it became harder for users to form

opinions for both real news and fake news). The tendencies were remarkable for fake news, indicating that people find it harder to perceive fake news as fake than to perceive real news as real. As shown in Fig.2, the average accuracy $\bar{A}$ was less than 80%, indicating that account imitation was always effective (i.e., account imitation did not fail, as mentioned in Section 5.2). The results for $T = 20$ hours showed the same tendencies as those for $T = 10$ hours (i.e., the average differences in accuracy $\Delta\bar{A}$, inaccuracy $\Delta\bar{I}$, and undetermined $\Delta\overline{UD}$ were $\Delta\bar{A} = 1.27\%$, $\Delta\bar{I} = 1.16\%$, and $\Delta\overline{UD} = 2.22\%$).

With account hijacking, accuracy and undetermined showed the same tendencies as account imitation for fake news. However, the results for real news were unexpected: accuracy was higher and undetermined was lower than with account imitation (i.e., green bars in accuracy for real news were higher and green bars in undetermined for real news were lower than those of the blue bars). This is because account hijacking sometimes fails due to the probability of success that depends on targets' security against hijacking, while account imitation succeeds regardless of their security against imitation.

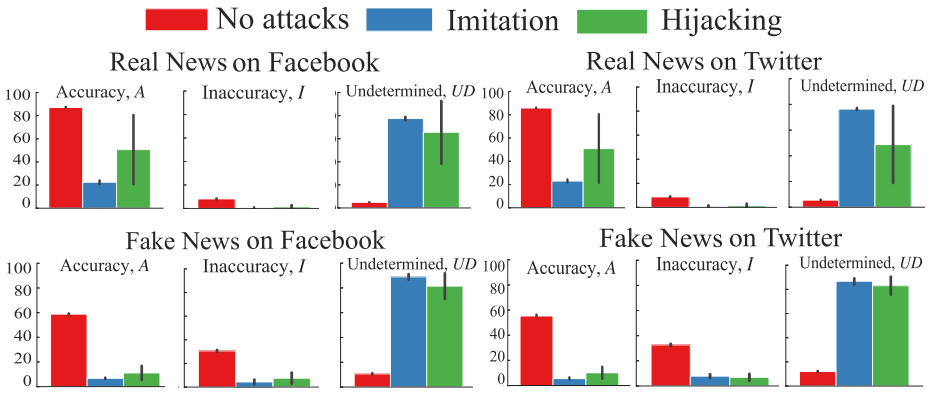

Fig. 2. Averages for accuracy, inaccuracy, and undetermined. "Imitation" and "Hijacking" results are for $T = 10$ hours; error bars are standard deviation.

- **Trust-Based Model**

  When there were no attacks, opinions gradually converged to 0 or 1, indicating that extremists bias users' opinions (which was also shown by Tsang and Larson [12]). Fig.3 shows an example of opinions converging to 0.

  When attackers whose opinions were 0 attacked the targets whose opinions were 1 ("Attacking 1-opinion targets" in Fig.3), opinions converged to 0 faster than when there were no attacks. This indicates that attackers facilitate the convergence of the opinion that benefits them. Fig.3 also shows that there were no substantial differences between account imitation and account hijacking. Furthermore, the results for $T = 20$ hours show the same tendencies as those for $T = 10$ hours. The average opinion update step at which simulations converged (i.e., all users did not update opinions more than a threshold ε) in 10 and 20 hours was 15.2 and 17.5, respectively. The same results were observed when attackers whose opinions were 1 attacked the

targets whose opinions were 0, which is not discussed due to space limitations (i.e., opinions converged to 1 more quickly than when there were no attacks, both with account imitation and account hijacking).
When attackers whose opinions were 1 attacked targets whose opinions were 0 ("Attacking 0-opinion targets" in Fig.3), the influence of opinion 0 was suppressed, and the opinions gradually converged to 1. This result corresponds to the situation in which attackers deter the influence of the targets' opinions and then facilitate the spread of opinions beneficial to the attackers. No substantial differences were found between account imitation and account hijacking. The average opinion update step at which simulations converged in 10 and 20 hours was 118.2 and 121.7, respectively. The same tendencies were observed when attackers whose opinions were 0 attacked the targets whose opinions were 1 (i.e., the influence of opinions 1 was suppressed, and the opinions gradually converged to 0).

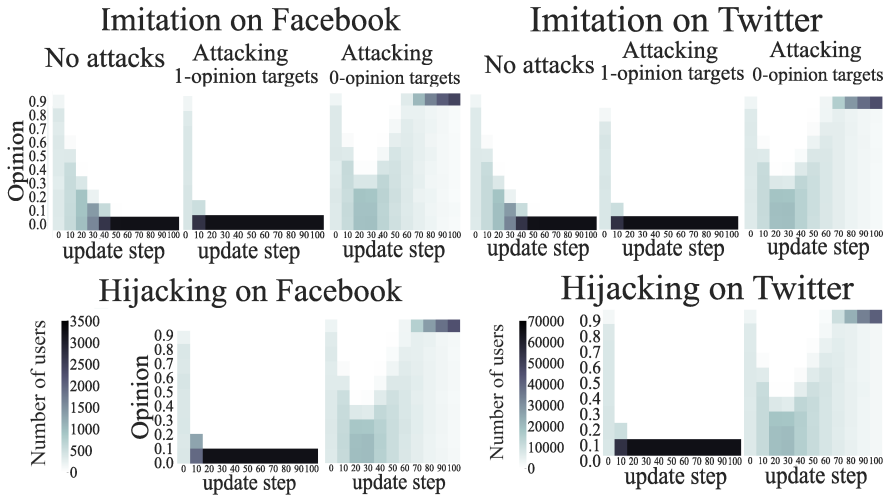

Fig. 3. Average opinion distribution for account imitation and account hijacking. "Attacking" results are for $T = 10$ hours.

6 Conclusion

We have described the dissemination of opinions regarding fake news by attackers on social networks. Previous models [4, 7] described the spread of correct (or incorrect) opinions assuming the presence of the ground truth while other models [3, 10, 12] described opinion convergence without assuming it. However, these models do not take attackers into account. Therefore, we modeled attackers who intentionally facilitate the spread of fake news by imitating or hijacking those who have many friends (i.e., influencers). We also incorporated attackers into two models, the AAT [7] and Tsang and Larson's Trust-Model [12].

Experimental results on real-world social networks revealed that, with the AAT-Based Model, account imitation was always effective (e.g., 1% of attackers

can deter 80.8% of the users from forming opinions) while account hijacking sometimes fails. With the Trust-Based Model, account imitation and account hijacking enabled 1% of attackers to quickly spread the opinions that supported attackers while suppressing the correct opinion convergence.

Our proposed models do not reflect psychological tendencies, such as the echo chamber effect and the backfire effect. Moreover, we have yet to develop effective countermeasures from our results. Future work will overcome these limitations.

References

1. Aaron, G., Chapin, L., Piscitello, D., Strutt, C.: Phishing landscape 2020 (2020)
2. Cui, Q., et al.: Tracking phishing attacks over time. In: Proceedings of the 26th International Conference on World Wide Web. pp. 667–676 (2017)
3. DeGroot, M.H.: Reaching a consensus. Journal of the American Statistical Association **69**(345), 118–121 (1974)
4. Glinton, R.T., Scerri, P., Sycara, K.: Towards the understanding of information dynamics in large scale networked systems. In: 2009 12th International Conference on Information Fusion. pp. 794–801. IEEE (2009)
5. Hameed, K., Rahman, N.: Today's social network sites: An analysis of emerging security risks and their counter measures. In: 2017 International Conference on Communication Technologies (ComTech). pp. 143–148. IEEE (2017)
6. Leskovec, J., Krevl, A.: SNAP Datasets: Stanford large network dataset collection. http://snap.stanford.edu/data (2014)
7. Pryymak, O., Rogers, A., Jennings, N.R.: Efficient opinion sharing in large decentralised teams. In: Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems. pp. 543–550 (2012)
8. Quattrociocchi, W., Scala, A., Sunstein, C.R.: Echo chambers on facebook. Available at SSRN 2795110 (2016)
9. Salahdine, F., Kaabouch, N.: Social engineering attacks: A survey. Future Internet **11**(4), 89 (2019)
10. Sasahara, K., et al.: Social influence and unfollowing accelerate the emergence of echo chambers. Journal of Computational Social Science (2020)
11. Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H.: Fake news detection on social media: A data mining perspective. ACM SIGKDD explorations newsletter **19**(1), 22–36 (2017)
12. Tsang, A., Larson, K.: Opinion dynamics of skeptical agents. In: Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems. pp. 277–284 (2014)
13. Vosoughi, S., Roy, D., Aral, S.: The spread of true and false news online. Science **359**(6380), 1146–1151 (2018)
14. Wang, Y., et al.: Eann: Event adversarial neural networks for multi-modal fake news detection. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 849–857 (2018)
15. Yang, S., et al.: Unsupervised fake news detection on social media: A generative approach. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 5644–5651 (2019)

Acknowledgements This work was supported by JSPS KAKENHI Grant Number JP21K11883.