Game Theoretic Opinion Models and Their Application in Processing Disinformation

Zhen Guo (presenter), Jin-Hee Cho

Computer Science @ Virginia Tech

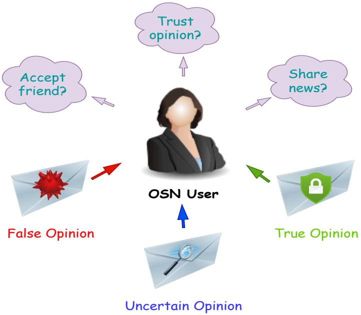
IEEE GLOBECOM 2021, Dec. 7–11, 2021

Outline

- **Motivation**
- **Key Contributions**
- **Related Work**
- **Opinion Models**
  - Five Types OMs
  - Interaction Model for Opinion Update
- **Game Theoretic Agent Models**
  - Attackers
  - Defender
  - Users
- **Experiment Setup**
- **Numerical Analysis & Results**
- **Conclusions**

Motivation

- Disinformation propagation becomes an urgent issue at all levels of our daily life, family and community via online social networks (OSNs).
- No prior work has analyzed opinion dynamics introduced by various game-theoretic opinion models and compared their abilities to handle disinformation.



- Subjective Logic (SL) can model a user's uncertain, subjective opinion under various types of opinion models.
- We take a game theoretic approach for an individual user to update his/her subjective opinion formulated by SL and investigate which opinion model(s) can better help combat disinformation.

Key Contributions

- Proposed a game theoretic opinion framework to handle disinformation with various types of opinion models.
- Formulated a user's uncertain opinion by a belief model, called *Subjective Logic* (SL), to analyze the dynamics of subjective and uncertain opinions.
- Designed attackers' various deception tactics of disinformation propagation by SL.
- Demonstrated optimal strategy choices by players along with their underlying reasons.
- Investigated how each opinion model contributes to combating disinformation propagation.

Related Work

- **Game theoretic information diffusion**
  - **Rumor spreading**: User's behaviors were modeled in evolutionary game theory (EGT) focusing on an attitude toward rumor propagation, severity of rumor or antirumor, or the anxiety level of society (Askarizadeh et al., 2019).
  - **Fake news spreading**: Users updated the reliability and doubt of their friends and then exchanged opinions by Bayesian estimation (Yoshikawa et al., 2020).
- **Opinion models by pairwise user interactions**
  - Subjective Logic proposed the consensus operator to update an opinion by estimating the trust in a friend's opinion (Jøsang, 2016).
  - An Assertion Model suggested the opinion update by a user's forgetfulnees, learning chance, and trust on a friend (Zinoviev & Duong, 2011).
  - Opinion exchange from social interactions was a herding behavior with collective friends (Sonowal et al., 2020).

Related Work

- **Game theoretic information diffusion**
  - **Rumor spreading**: User's behaviors were modeled in evolutionary game theory (EGT) focusing on an attitude toward rumor propagation, severity of rumor or antirumor, or the anxiety level of society (Askarizadeh et al., 2019).
  - **Fake news spreading**: Users updated the reliability and doubt of their friends and then exchanged opinions by Bayesian estimation (Yoshikawa et al., 2020).
- **Opinion models by pairwise user interactions**
  - Subjective Logic proposed the consensus operator to update an opinion by estimating the trust in a friend's opinion (Jøsang, 2016).
  - An Assertion Model suggested the opinion update by a user's forgetfulnees, learning chance, and trust on a friend (Zinoviev & Duong, 2011).
  - Opinion exchange from social interactions was a herding behavior with collective friends (Sonowal et al., 2020).

Related Work

- **Game theoretic information diffusion**
  - **Rumor spreading**: User's behaviors were modeled in evolutionary game theory (EGT) focusing on an attitude toward rumor propagation,

Limitation

- **There has been little work to investigate what types of interaction can better combat disinformation propagation.**
- **Little research has examined the impact of existing opinion models on mitigating disinformation propagation.**

forgetfulnees, learning chance, and trust on a friend (Zinoviev & Duong, 2011).

  - Opinion exchange from social interactions was a herding behavior with collective friends (Sonowal et al., 2020).

Opinion Models

Opinion Formation Using Subjective Logic (SL)

- A binomial opinion for a given proposition is $\omega = \{b, d, u, a\}$ for b (belief), d (disbelief), u (uncertainty), and a (base rate), where $b + d + u = 1$.
- Expected probability of belief and disbelief by agent i is $P(b_i) = b_i + a_i u_i$ and $P(d_i) = d_i + (1 - a_i)u_i$, where $P(b_i) + P(d_i) = 1$

Initialization of Opinions

- True opinion: $\{b \to 1; d \to 0; u \to 0; a = 1\}$, implying highly true.
- False opinion: $\{b \to 0; d \to 1; u \to 0; a = 0\}$, implying highly false.
- Uncertain opinion: $\{b \to 0; d \to 0; u \to 1; a = 0.5\}$, implying highly uncertain without strong preference for the rest of opinions.

Opinion Update

- Stubborn zealots (e.g., true or false informers) propagate true or false opinions but refuse to update their extreme opinions.
- Other users update opinions through pairwise interactions by the five types of opinion models (OM).

Opinion Models

Consensus Operator

- User i update opinion ω_i with ω_j by the **consensus** operator $\oplus$:

$$b_i \oplus b_{i\otimes j} = \frac{b_i(1-c_i^j(1-u_j))+c_i^j b_j u_i}{\beta}, d_i \oplus d_{i\otimes j} = \frac{d_i(1-c_i^j(1-u_j))+c_i^j d_j u_i}{\beta},$$

$$u_i \oplus u_{i\otimes j} = \frac{u_i(1-c_i^j(1-u_j))}{\beta}, a_i \oplus a_{i\otimes j} = \frac{(a_i-(a_i+a_j)u_i)(1-c_i^j(1-u_j))+a_j u_i}{\beta-u_i(1-c_i^j(1-u_j))}.$$

Discounting Operator

- Depending on an OM, user i can estimate trust of encountered user j's opinion and use it as a filter, c_i^j, based on the discounting operator $\otimes$.

$$b_{i\otimes j} = c_i^j b_j, \quad d_{i\otimes j} = c_i^j d_j,$$
$$u_{i\otimes j} = 1 - b_{i\otimes j} - d_{i\otimes j}, \quad a_{i\otimes j} = a_j.$$

Five Types of Opinion Models (OMs)

- **Uncertainty-based OM**
  - Perceived uncertainty towards an encountered user's opinion (Cho, 2018).
  - Uncertainty maximization: convert conflicting evidence to vacuity (Jøsang, 2016).
  - Uncertainty-based discounting filer for $\omega_{i \otimes j}$: $uc_i^j = (1 - u_i)(1 - u_j)$
- **Homophily-based OM**
  - Like-minded users estimate the extent of homophily (Li et al., 2015).
  - Discounting filer hc_i^j is measured by cosine similarity of two users' opinions (i.e., belief and disbelief).
- **Encounter-based OM**
  - Baseline model to update user i's opinion on the consensus operator $\omega_i \oplus \omega_i$ without any filter (i.e., $c_i^j = 1$) (Jøsang, 2016).

Five Types of Opinion Models (OMs) – (cont'd)

Assertion-based OM

- An assertion, A_i, with the quantity of knowledge and subjective prior belief by $A_i = \{k_i, \mathrm{spb}_i\}$ (Zinoviev & Duong, 2011).
- Translated to SL opinion ω_i by:

$$k_{i \oplus j} = k_i + k_j(1 - k_i),\ a_{i \otimes j} = a_i + b_j a_j(1 - a_i),\ \text{For } k \in [b, d, u].$$

Herding-based OM

- Updating one's opinion based on all his/her neighbors' opinions (Sonowal et al., 2020).
- Transformed the original model to the SL structure by:

$$x_i = \min[1, x_i + \frac{u_i}{|F_i|} \sum_{j \in F_i} (1 - u_j)(x_j - x_i)],$$

$$u_i = 1 - (b_i + d_i),\ \text{For } k \in [b, d, a].$$

Interaction Model for Opinion Update

- **Sharing**
  - **Pair-wise interaction** (P_i^f): interacting with other users by leaving comments or receiving messages
  - **Posting** (P_i^p): sharing opinion with all friends by posting
  - **The probability of user i selecting j to interact**: $P_{ij} = \frac{P_j^f + P_j^p}{\sum_{k \in F_i}(P_k^f + P_k^p)}$
- **Friending and unfriending**
  - Projected difference between two opinions: $PD_{ij} = \frac{|b_i - b_j| + |d_i - d_j|}{2}$
  - **Friending**: Each user will invite a friend by the Price Model. In Uncertainty-based OM, user j only accepts a friend request from user i when $u_i < \phi_j$. Other OM users accept if $PD_{ji} < \phi_j$. Reject the request, otherwise.
  - **Unfriending**: User i in uncertainty-based OM will quit the relationship with current friend j if $\phi_i < u_j < 0.5$. Other OMs user i will unfriend with user j if $PD_{ji} > \phi_j$.

Game Theoretic Agent Models - Attackers

- **Attacker strategies**
  - Aim to maximize disinformation propagation by deception tactics (Kopp et al., 2018).
  - **Degradation (DG)**: increase confusion of users by propagating highly uncertain opinions
  - **Corruption (C)**: mislead users by providing disinformation such as an opposed opinion by switching b and d.
  - **Denial (DN)**: block the access to true information by dropping true information propagation.
  - **Subversion (S)**: increase the amount of disinformation propagated
- **Expected payoffs**

$$EP_k^{A_i}(a^U, a^D) = \sum_{\ell \in D} \sum_{m \in U} p_\ell^D \cdot p_m^U \cdot u_{k\ell m}^{ij},$$

  - $u_{k\ell m}^{ij}$ is the utility of attacker's strategy k: $u_{k\ell m}^{ij} = ds(k, m, \omega_i, \omega_j) - g_\ell$.
  - $ds(k, m, \omega_i, \omega_j)$ is the gain of similarity to the false opinion if k is taken.
  - g_ℓ is the attacker's loss when the defender takes ℓ strategy: similarity between the true opinion and each of all users' opinion.

Game Theoretic Agent Models - Defender

Defender strategies

- A defender, as a system administrator, aims to protect a given OSN platform.
- **Terminating a malicious user (T)**: detect a user with malicious intent based on the amount (N_R) of misconduct reports.
- **Monitoring a suspect user (M)**: increase the defender's confidence in detecting a correct malicious user by monitoring further.

Expected payoffs

$$EP_{\ell}^{D}(a^A) = \sum_{k \in A} p_k^A \cdot u_{\ell k}^{D},$$

- $u_{\ell k}^{D}$ is the utility of defense strategy ℓ under attack strategy k: $u_{\ell k}^{D} = s(\ell, k, \omega_T, \omega') - c_{\ell}$
- $s(\ell, k, \omega_T, \omega')$ is the degree of the mean similarity between the truth opinion and the expected opinion ω' of all users.
- c_{ℓ} is a constant cost incurred by taking defense strategy ℓ, where the cost of taking T and M is set to 0.1 and 0, respectively.

Game Theoretic Agent Models - User

- **User strategies**
  - Five user types based on the corresponding opinion model, by types based on: uncertainty (U), homophily (H), Encounter (E), Assertion (A), and Herding (HE).
  - **Updating and sharing (SU)**: update an opinion and share the opinion.
  - **Updating (U)**: update the current opinion but not to share it.
  - **No Updating (NU)**: refuse to update the current opinion after interaction.
- **Expected payoffs**
  - User i interacts his/her opinion with either a legitimate user or an attacker to calculate the expected payoffs by:

$$EP_m^{U_i}(a^{U_j}) = p_{U_i}^{A_j} \cdot u_m^{U_i A_j} + (1 - p_{U_i}^{A_j}) \cdot u_m^{U_i U_j}.$$

  - $p_{U_i}^{A_j}$ refers to the probability of user j being an attacker.

Game Theoretic Agent Models - User (cont'd)

- **Expected payoffs (cont'd)**
  - The utility of strategy m taken by user i when encountering an attacker or a legitimate user are $u_m^{U_i A_j}$ and $u_m^{U_i U_j}$:

$$\text{meeting an attacker:} \quad u_m^{U_i A_j} = \begin{cases} \sum_{k \in A} p_k^{A_i A_j} \cdot -s(m, \omega_F, \omega_i, \omega_j) & \text{if } m \neq a_3^U; \\ 0 & \text{if } m = a_3^U, \end{cases}$$

$$\text{meeting a user:} \quad u_m^{U_i U_j} = \begin{cases} \sum_{m' \in \mathcal{U}_j} p_{m'}^{U_j} \cdot uc_{im'}^{j} & \text{if } j \text{ is a U-type user;} \\ \sum_{m' \in \mathcal{U}_j} p_{m'}^{U_j} \cdot hc_{im'}^{j} & \text{otherwise,} \end{cases}$$

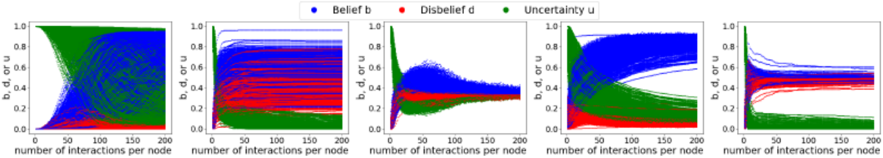
Experiment Setup

- **Dataset**
  - A sample of 1,000 users from a real Twitter dataset 1KS-10KN (Yang et al., 2012).
- **Metrics**
  - Opinions of agents ($\omega_i = \{b_i, d_i, u_i, a_i\}$): Four opinion masses of the SL-based opinion for user i and the expected belief $P(b_i)$.
  - Probability distribution of taking strategies: The best chosen strategies for each player type.

KEY PARAMETERS AND DEFAULT VALUES

Param.	Default value	Param.	Default value
T	200	P_i^f	0.142 (mean)
N	1,000	P_i^p	0.186 (mean)
P_{true}	0.1	c_ℓ	T: 0.1, M: 0
P_{false}	0.1	ρ	*Normal* (0.5, 0.05)
ϕ	*Normal* (0.1, 0.1)	N_R	3
ξ	0.05	$p_{U_i}^{A_j}$	0.1

Analysis of Opinions Dynamics in the Five OMs



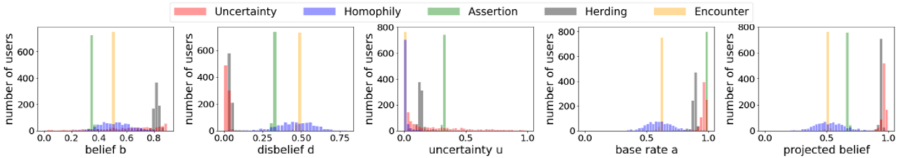
(a) Uncertainty-based OM (b) Homophily-based OM (c) Assertion-based OM (d) Herding-based OM (e) Encounter-based OM

Fig. 1: The opinions dynamics over 200 user interactions by the scatter plots of all individual opinions in terms of belief (b), disbelief (d) and uncertainty (u). The five subfigures have the same configurations except for the OM.

Users updating opinions under different OMs show distinctive dynamics and U type users can reach the highest beliefs, followed by HE type users.

- **U type users**: More users believe true information as interacting more because the U type users rely on perceived uncertainty for opinion update.
- **HE type users**: High belief values but the uncertainty in their opinions is much lower than that of U type users.
- **H type users**: Diverse trends of believing both true and false information are observed.

Analysis of Uncertain Opinions in the Five OMs



(a) B (b) D (c) U (d) A (e) $P(b)$

Fig. 2: The histograms of B, D, U and A and projected belief $P(b)$ from five types of opinion update models.

The opinions of U type users show high belief, lowest disbelief, and a higher base rate evolved from initial base rate 0.5.

- The projected belief (commonly used for actual making decision) by U type users is higher than that of HE type users.
- The effective defense of U type users against disinformation removes connections with false informers having uncertain and noisy opinions.
- H type users trust false information more since they cannot identify noisy and uncertain opinions.

Analysis of Strategy Selection

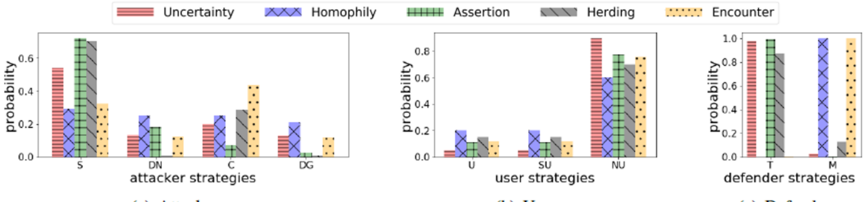


Fig. 3: The distributions of the strategies taken by three types of players during 200 interactions.

- **Attackers**: Adopt S and C more in the U type network but choose all strategies equally in the H type network.
- **Users**: All take strategy NU much more than U and SU while U type users have the least chance to update while H type users update more frequently.
- **Defender**: U, A and HE type networks prefer T for over 80% times. But H and E type networks barely take T to remove a malicious account.

Key Findings

1. Uncertainty and herding-based opinion models (OMs) can more effectively help users resist disinformation propagation and believe true information.
2. On the other hand, homophily, assertion, and encounter-based OMs can easily deceive users and accept disinformation.
3. Users using uncertainty-based OM perform the best among all other comparing OMs in effectively mitigating the impact of disinformation propagation.
4. Uncertainty-based opinion model showed outperformance because it can substantially minimize uncertainty and adjust prior beliefs towards true information, guiding people to filter out disinformation.

Any Questions?

Thank You!

Zhen Guo
zguo@vt.edu

List of Notations of Design Parameters

Notation	Description
ω, ω_i	User i's SL-based binary opinion
P_{true}	Proportion of true informers
P_{false}	Proportion of false informers
P_{uc}	Proportion of normal users
$\oplus$	Consensus operator from SL
$\otimes$	Trust operator from SL
c_i^j	Discounting operator from SL
ξ	Threshold of uncertainty maximization
P_i^f	User i's feeding probability
P_i^p	User i's posting probability
PD_{ij}	Projected discrepancy between two opinions
ϕ	Threshold to accept or request a friend
ω_F	False opinion $\{0, 1, 0, 0\}$
ω_T	True opinion $\{1, 0, 0, 1\}$
$p_{U_i}^{A_j}$	Probability of user j as an attacker
N_R	Number of reports to alert a defender
ρ	Tolerance to report a malicious user

Experiment Setup

- In the first interaction, all players will start by:
  - **1.A**: Attackers play a move of Subversion (S) to spread false opinions;
  - **1.B**: A user i chooses an existing friend j by j's sharing probability P_{ij}. Both of them decide their strategies and update opinions as needed;
  - **1.C**: If attacker j is selected by legitimate user i who trusts attacker j's opinion, attacker j will also share with other friend ω_i received from user i and spread the fabricated opinion to mislead other users;
  - **1.D**: The defender will take a defense strategy on a suspect user based on the minimum number of reports, N_R;
  - **1.E**: Each user will follow the descriptions of the friending and unfriending procedures.
- Starting from the second interaction to the T-th interaction of the repeated game, each player will keep performing 1.C, 1.D, 1.B and 1.E, accordingly but chooses a best strategy based on the respective utility.