Uncertainty-Aware Opinion Inference Under Adversarial Attacks

Adil Alim[1], Xujiang Zhao[2], Jin-Hee Cho[3],Feng Chen[2]

[1]*Department of Computer Science*
University at Albany - SUNY
Albany, NY, USA
aalimu@albany.edu

[2]*Department of Computer Science*
The University of Texas at Dallas
Dallas, Texas, USA
{Xujiang.Zhao, Feng.Chen}@utdallas.edu

[3]*Department of Computer Science*
Virginia Tech
Falls Church, VA, USA
jicho@vt.edu

***Abstract*—Inference of unknown opinions with uncertain, adversarial (e.g., incorrect or conflicting) evidence in large datasets is not a trivial task. Without proper handling, it can easily mislead decision making in data mining tasks. In this work, we propose a highly scalable opinion inference probabilistic model, namely Adversarial Collective Opinion Inference (Adv-COI), which provides a solution to infer unknown opinions with high scalability and robustness under the presence of uncertain, adversarial evidence by enhancing Collective Subjective Logic (CSL) which is developed by combining SL and Probabilistic Soft Logic (PSL). The key idea behind the Adv-COI is to learn a model of robust ways against uncertain, adversarial evidence which is formulated as a min-max problem. We validate the outperformance of the Adv-COI compared to baseline models and its competitive counterparts under possible adversarial attacks on the logic-rule based structured data and white and black box adversarial attacks under both clean and perturbed semi-synthetic and real-world datasets in three real world applications. The results show that the Adv-COI generates the lowest mean absolute error in the expected truth probability while producing the lowest running time among all.**

I. Introduction

Under highly dynamic communication networks, information received for decision making may be uncertain, incomplete, modified/forged and/or missing due to unreliable medium and/or the presence of malicious entities. Research on decision making under uncertainty has been studied in evidence or belief theories considering uncertainty reasoning. On the other hand, the data mining research mainly focused on efficient reasoning of data uncertainty that impacts ways of providing solutions of other data mining tasks.

In the belief or evidence model research (i.e., knowledge representation and reasoning, or KRR), Subjective Logic (SL) [7] has been proposed to explicitly deal with uncertainty in subjective opinions. SL is a probabilistic logic representing an opinion in terms of belief, disbelief, and uncertainty for a binary opinion (i.e., pro vs. con). For multinomial or hypernomial opinions, SL provides a set of logic operators that allow deriving structural relations between opinions (i.e., random variables) in a network with entities for vertices and their relationships for edges. However, the dyadic combinations of different opinions in SL are well-known as its limitation in scalability [12]. Probabilistic Soft Logic (PSL) [3] provides collective reasoning with high scalability on the relationships between opinions based on the known truth probability; but it cannot deal with uncertainty. [12] proposed Collective SL (CSL) by combining the merits of both SL and PSL capable of handling uncertain opinions as well as processing large-scale network data. However, CSL is not robust to simple noisy or adversarial perturbed data, which is confirmed in our experimental findings.

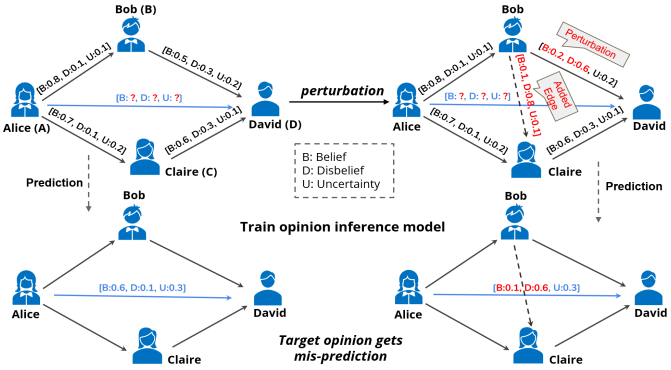


Fig. 1: **An illustration of adversarial attacks** on the logic rule-based structural data (Trust inference), given opinions of relationships in (A, B), (B, D), (A, C), (C, D). Our goal is to predict trust opinion from A to D. After the attack on both *opinion* and *graph structure*, A has a high chance to distrust D, unlike the results on the clean data.

The key research questions we aim to answer are: **How to develop a model that can efficiently fool uncertainty in inferring unknown opinions in large network data? How reliable are their results?** To the best of our knowledge, these research questions have not been answered in both belief/evidence theory and data mining research.

In this paper, we design attacks injecting adversarial evidence to test the robustness of uncertainty-based opinion prediction probabilistic models, where the data structures are represented by first-order logic rules. In addition, we investigate the robustness and scalability of our Adv-COI model for accurate opinion inference optimization under the attacks. Adv-COI collectively reasons unknown opinions using the relations represented by the logical rules between given/unknown opinions, as demonstrated in Fig. 1.

We formulate this opinion prediction problem as a min-

max problem to measure the robustness against adversarial evidence (e.g., conflicting or wrong evidence). As a "black-box," inference method, Adv-COI infers adversarial training opinions (i.e., attacker's opinions) without knowing the ground truth of testing opinions (i.e., target opinions). Attacking a model tends to be easier than defending against attacks. We proposed novel augmented first order rules while inferring unknown opinions, in which Adv-COI learns a soft indicator for each training opinion to decrease the influence of perturbed adversarial training opinions. We validate the performance of the Adv-COI in terms of the opinion prediction accuracy, robustness against adversarial evidence, and scalability by comparing it with baseline and other competitive counterparts based on the semi-synthetic and real world datasets. The **key contributions** of this work are:

- **A novel, robust opinion inference model** is proposed to predict unknown opinions under attack injecting adversarial evidence based on the structured data, which has not been considered in the literature. We introduce a projected gradient based opinion perturbation attack, and a greedy algorithm for the structure attack on network graph, and a defense mechanism against them.
- **The proposed scalable opinion inference algorithm**, Adv-COI, is highly scalable in predicting unknown opinions in large, adversarial network data, by benefiting from learning and defending against opinion and structure perturbation attacks while inferring the unknown opinions.
- **The performance of the Adv-COI is validated** based on four semi-synthetic and two real world datasets showing its outperformance over other counterparts under different level of the white and black box attacks (opinion and structure unnoticeable perturbation attacks) in three real-world applications.

II. Related Work

Probabilistic and Belief Models: Due to lack of evidence or knowledge, a large volume of works has been proposed to model uncertainty in network data as a joint distribution over a set of variables, in which each variable relates to a node in the network. To address the computational limitations of Markov Logic Networks (MLNs) and Markov Random Fields (MRFs), a new probabilistic logic, called PSL [5], is designed to define relations between the truth probabilities of binary variables, and the inference of PSL rules based on Hinge-Loss-MRFs [3]. However, PSL has not explicitly dealt with uncertainty in derived relations of the truth probabilities. In evidence / belief theory, SL is proposed to define an opinion explicitly dealing with uncertainty. SL offers a variety of operators to fuse multiple opinions. But SL is limited in its scalability as it combines opinions in a dyadic manner.

Hybrid or deep learning (DL)-based uncertainty models: CSL [12] reasons an opinion under uncertainty (i.e., vacuity) for a scenario where all the node-level opinions in a network have the same uncertainties but different belief or disbelief. Similar to CSL, Adv-COI combines the merits of SL and PSL; however, CSL lacks tolerance or resistance against noisy or perturbed data. GCN-VAE-opinion (graphical convolutional network-variational autoencoder-opinion) [15] is one of DL-based opinion inference models adopting GCN and VAE to deal with uncertain opinions characterized by a set of heterogeneous belief and uncertainty in a network data.

Adversarial Attack: Recently, researchers have studied the vulnerability of adversarial machine learning (ML). Most approaches have focused on DL models, studying the effect of adversarial ML on image classification, Neural Network (NN) policies [21], and autonomous driving system, speech recognition, or text classification models. [22], [23] proposed adversarial attack methods on DL based graph learning tasks. Several studies investigated how to generate the adversarial examples [17]–[19].

Unlike the above existing state-of-the-art approaches, our work studies uncertain, subjective opinions in large network data where there exists adversarial evidence whose injection is first designed to test the robustness of the proposed Adv-COI.

III. Background

A. Subjective Logic

In SL, a binomial opinion about the truth of a proposition x is represented as the tuple (b_x, d_x, u_x, a_x), where b_x is the belief that is true, d_x is the belief that x is false, u_x is the uncertainty, and a_x is the base rate (a prior probability in the absence of evidence), as well as $b_x + d_x + u_x = 1$ and $b_x, d_x, u_x, a_x \in [0, 1]$. A Beta PDF is the same as a binomial opinion through a specific bijective mapping [7]. Given the binary domain $\mathbb{X} = \{x, \bar{x}\}$ and the value $x \in \mathbb{X}$, Beta(p_x) is the probability density function $\text{Beta}(p_x; \alpha, \beta) : [0, 1] \rightarrow \mathbb{R}_{\geq 0}$ where $p_x + p_{\bar{x}} = 1$. The Beta PDF is given by:

$$\textbf{Beta}(p_x; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) + \Gamma(\beta)} (p_x)^{\alpha-1} (1 - p_x)^{\beta-1}, \tag{1}$$

where $\alpha, \beta > 0$. The α and β parameters can simply be represented by the base rate a_x and the observation evidence (r_x, s_x) where r_x is the amount of positive evidence and s_x is the amount of negative evidence $\alpha = r_x + a_x W$, $\beta = s_x + (1 - a_x)W$, W is the non-informative prior weight in the absence of r_x or s_x. The expected probability of the Beta PDF is:

$$\mathbb{E}(x) = \frac{\alpha}{\alpha + \beta} = \frac{r_x + a_x W}{r_x + s_x + W}. \tag{2}$$

The equivalence of a binomial opinion and a Beta PDF can be achieved through the following mapping rule:

$$b_x = \frac{r_x}{r_x + s_x + W}, d_x = \frac{s_x}{r_x + s_x + W}, u_x = \frac{W}{r_x + s_x + W}. \tag{3}$$

B. Adversarial Attack and Defense

1) Adversarial Attack: We introduce two kinds of adversarial attacks related to our work: feature and structure attacks.

TABLE I: Key notations

$\mathbf{X_S} = [X_{S_1}, \ldots, X_{S_N}]$, $\mathbf{p}_S = [\mathbf{p}_{S_1}, \ldots, \mathbf{p}_{S_N}]$, $\omega_S = [\omega_{S_1}, \ldots, \omega_{S_N}]$	$\mathbf{X_S}$ is a vector of N input binary random variables whose subjective opinions are unknown. $\mathbf{p}_S$ and ω_S are the corresponding vectors of truth probabilities and subjective opinions of $\mathbf{X_S}$, respectively.
$\mathbf{X}_{\overline{S}} = [X_{\overline{S}_1}, \ldots, X_{\overline{S}_M}]$, $\mathbf{p}_{\overline{S}} = [\mathbf{p}_{\overline{S}_1}, \ldots, \mathbf{p}_{\overline{S}_M}]$, $\omega_{\overline{S}} = [\omega_{\overline{S}_1}, \ldots, \omega_{\overline{S}_M}]$	$\mathbf{X}_{\overline{S}}$ is a vector of M input binary random variables whose subjective opinions are known. $\mathbf{p}_{\overline{S}}$ and $\omega_{\overline{S}}$ are the corresponding vectors of truth probabilities and subjective opinions of $\mathbf{X}_{\overline{S}}$, respectively.
S_i (or $\overline{S}_i$)	indicates the index of i^{th} element in S (or $\overline{S}_i$)
$\omega_{S_i} = (b_{S_i}, d_{S_i}, u_{S_i}, a_{S_i})$	A binomial subjective opinion of a binary random variable X_{S_i} as defined in Section III-A
$\mathbf{b}_{\overline{S}} = (b_{\overline{S}_1}, \cdots, b_{\overline{S}_M})$	a vector of M binary random variables which we want to learn during our adversarial inference.
$q(\mathbf{p}, \mathbf{b}_{\overline{S}})$	A new pdf function that fits the logic rules defined in $\mathcal{R}$ as well as meets the minimal KL-divergence distance to the posterior Prob$(\mathbf{p}, \mathbf{b}_{\overline{S}} \vert \mathbf{X}_{\overline{S}}; \boldsymbol{\omega}_S, \theta)$ (See Eq. (17)).

a) Projected Gradient Descent (PGD) attack: PGD attack [24] is an iterative variant of Fast Gradient Sign Method (FGSM) [20], in each iteration, PGD follows the update rule:

$$x_{l+1}^{adv} = \mathbf{clip}_{[a,b]}\{x_l + \epsilon \cdot \text{sign}(\nabla_{x_l^{adv}} \mathcal{L}(x_l^{adv}, y, \theta))\} \tag{4}$$

where the outer clip function $\mathbf{clip}_{[a,b]}(\cdot)$ keeps x_{l+1}^{adv} within a predefined perturbation range. PGD can also be interpreted as an iterative algorithm to solve the following problem:

$$\max_{x^{adv}: \|x^{adv} - x\|_C \leq \gamma} \mathcal{L}(x^{adv}, y; \theta) \tag{5}$$

where $C \in \{0, 2, \infty\}$, γ represents the perturbed level. For example, ℓ_∞-norm distance measures **the maximum change to any of the coordinates**.

b) Attack on graph structure: [22], [23] considered adversarial attacks on graph data focusing on specific types of attacks on DL models. Structure perturbation includes adding/removing edges/nodes aiming to increase the misclassification of the target nodes/edges. Each work formulated its own unified form by:

$$\max_{\mathcal{G}', \mathcal{D}(\mathcal{G}, \mathcal{G}') \leq \Delta} \sum_i \mathcal{L}(f_{\theta^*}(\mathcal{G}'_{c_i}, c_i), y) \tag{6}$$

where $\mathcal{L}$ is the loss function, $\mathcal{G}'$ final perturbed graph, c_i may be a target node/edge and $\mathcal{G}'_{c_i}$ its associated perturbed graph, $\mathcal{D}(\cdot)$ is a distance metric, and Δ is the perturbed cost/budget.

2) Adversarial Training: Adversarial training is a defense method against adversarial samples in [20]. This approach attempts to improve the robustness of the model by training or inferring it together with adversarial samples. Adversarial training solves the following min-max problem:

$$\min_\theta \max_{x^{adv}: \mathcal{D}(x, x^{adv}) \leq \gamma} \mathcal{L}(x^{adv}, y, \theta), \tag{7}$$

where $\mathcal{D}(x, x^{adv})$ represents certain distance metric between x and x^{adv}. The **inner maximization** problem is equivalent to constructing correct or strongest adversarial samples. If ℓ_∞ distance is employed as the distance metric $\mathcal{D}(x, x^{adv})$, the inner maximization problem is equivalent to the adversarial problem solved by PGD, i.e., Eq. (5). The **outer minimization** is the standard training or inference procedure.

IV. Adv-COI probabilistic inference model

We introduce our proposed method Adv-COI, a probabilistic inference model for predicting unknown opinions on the structured data where the structural relationships between data are captured by first-order logic rules. We study the adversarial attacks to the perturbation of given opinions or graph structure, and proposed a novel defense mechanism against adversarial attack. We proposed a novel augmented first order logic rules to enhance the robustness of our model against adversarial attacks. Our Adv-COI mainly is two-fold: The ***generation of adversarial opinions*** and the ***robust opinion inference***.

A. Problem Formulation

We consider the task of **unknown opinion prediction against adversarial attacks.** To formally put, given a factor graph network, defined as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, and $\mathcal{V}$ and $\mathcal{E}$ are the node and edge set of the graph $\mathcal{G}$, $X \in \{0, 1\}^{|\mathcal{E}| \times T}$ represents the edges associated observations. Without loss of generality, we assume the edge-ids to be $\mathcal{E}_I = \{1, \cdots, N + M\}$, and the node-ids $\mathcal{V}_I = \{1, \cdots, |N_\mathcal{V}|\}$, $N_\mathcal{V}$ is graph size. Each edge index of $i \in \mathcal{E}_I$ is associated with a set of local observation X_i. Assume we are only given the observation of a subset of edges $\overline{S} \in \mathcal{E}_I, |\overline{S}| = M$, and the set of local observations for edges in $\overline{S}$ is denoted by $\mathbf{X}_{\overline{\mathbf{S}}} = \{X_i | i \in \overline{S}\}$. The opinions of these variables are estimated based on the given observations, and denoted by $\omega_{\overline{S}} = (\omega_{\overline{S}_1}, \cdots, \omega_{\overline{S}_M})$, implying that the PDF of the truth probability of the variable $X_{\overline{S}_i}$, $p_{\overline{S}_i}$, is Beta$(p_{\overline{S}_i}; \omega_{\overline{S}_i})$, where $\overline{S}_i$ indicates the index of i^{th} element in $\overline{S}$.

Given the above, we aim to predict the **unknown opinions** of the rest of the variables (**target opinion**), $S = \mathcal{E}_I \setminus \overline{S}, |S| = N$. We denote the observation by $\mathbf{X_S} = \{X_i | i \in S\}$, and the unknown opinions as $\omega_S = (\omega_{S_1}, \cdots, \omega_{S_N})$, implying that the PDF of the truth probability of the variable, X_{S_i}, p_{S_i}, is Beta$(p_{S_i}; \omega_{S_i})$. We denote $\mathbf{p}_S = (p_{S_1}, \cdots, p_{S_N})$, $\mathbf{p}_{\overline{S}} = (p_{\overline{S}_1}, \cdots, p_{\overline{S}_M})$, and assume there is an operator Γ hat concatenates $\mathbf{p}_S$ and $\mathbf{p}_{\overline{S}}$ in the correct order, $\mathbf{p} = \Gamma(\mathbf{p}_S, \mathbf{p}_{\overline{S}})$. $\mathbf{b}_{\overline{S}} = (b_{\overline{S}_1}, \cdots, b_{\overline{S}_M})$ is a vector of M binary random variables which we want to learn during our adversarial inference.

Given

- The observations $\mathbf{X}_{\overline{\mathbf{S}}}$,and the opinions $\omega_{\overline{S}} = (\omega_{\overline{S}_1}, \cdots, \omega_{\overline{S}_M})$ of the set of variables in $\overline{S}$, where $\omega_{\overline{S}_i} = (\alpha_{\overline{S}_i}, \beta_{\overline{S}_i})$,
- $\mathcal{R} = \{r_k, \rho_k\}_{k=1}^K$, a set of logic rules, in which r_k and ρ_k is the k-th rule and its weight. An **augmented logic rule** r_k is defined by:

$$r_k := \bigwedge_{i \in I^-_{S,k}} p_{S_i} \bigwedge_{i \in I^-_{\overline{S},k}} \left(p_{\overline{S}_i} \wedge (1 - b_{\overline{S}_i})\right) \rightarrow \bigvee_{i \in I^+_{S,k}} p_{S_i} \bigvee_{i \in I^+_{\overline{S},k}} \left(p_{\overline{S}_i} \vee (1 - b_{\overline{S}_i})\right), \tag{8}$$

where $I^-_{S,k}$ and $I^-_{\overline{S},k}$ refer to the indices of variables in S and $\overline{S}$ that appear in the head of the logic rule r_k, respectively; and $I^+_{S,k}$ and $I^+_{\overline{S},k}$ refer to the indices of variables in S and $\overline{S}$ that appear in the tail of the logic rule r_k. Without considering the weight of the rule, the distance to meeting rule r_k is $d_k(\mathbf{p}, \mathbf{b}_{\overline{S}}) = \max\left\{\ell_k(\mathbf{p}, \mathbf{b}_{\overline{S}}), 0\right\}$, where:

$$\ell_k(\mathbf{p}, \mathbf{b}_{\overline{S}}) = 1 - \sum_{i \in I^+_{S,k}} p_{S_i} - \sum_{i \in I^+_{\overline{S},k}} (p_{\overline{S}_i} + 1 - b_{\overline{S}_i}) - \sum_{i \in I^-_{S,k}} (1 - p_{S_i}) - \sum_{i \in I^-_{\overline{S},k}} (b_{\overline{S}_i} - p_{\overline{S}_i}). \tag{9}$$

Goal: Predict ω_S, the **target** opinion of variables in S.

B. Adversarial Attack on Opinion Network

This work focuses on **feature** and **structure** adversarial attacks on graph structured data.

1) Feature Perturbation: is an indirect attack on target opinions. We formulate the problem as follows:

$$\max_{\mathcal{D}(\cdot,\cdot)\leq\gamma} \mathcal{L}(\cdot) \tag{10}$$

where $\mathcal{L}(\cdot)$ is an objective, $\mathcal{D}$ is widely used distance metrices, l_0, l_2 or l_∞, with γ being the constraint perturbation level.

2) Structure Perturbation: is a structure attack on factor graph $\mathcal{G}$, $\mathcal{D}(\cdot) \leq \Delta$ represents a measure of 'closeness' for (attributed) graphs, and k is a parameter denoting the distance/cost budget for the total graph perturbation. In this setting, we formulate the problem as:

$$\max_{\mathcal{G}', \mathcal{D}(\mathcal{G},\mathcal{G}')\leq\Delta} \mathcal{L}_{\mathcal{G}'}(\cdot), \tag{11}$$

where $\mathcal{L}_{\mathcal{G}'}$ (or $\mathcal{L}$) is our objective, detailed in the following sections, associated with the perturbed factor graph $\mathcal{G}'$. This problem is intractable to solve exactly due to discrete domain and constraints. Hence, we propose a scalable greedy approximation algorithm to solve the optimization problem in Eq. (11), where $\mathcal{G}'$ is obtained after adding/removing edges on the original graph $\mathcal{G}$.

For a targeted edge c_0, we manipulate (add/remove) the related candidate edges $\{c_1, \cdots, c_m\}$ with a budget constrain ($\mathcal{D}(\mathcal{G}, \mathcal{G}') \leq \Delta$). Under the budget constrain, for a given set of target edges, the modification of $\mathcal{G}$ is performed by sequentially modifying edges of $\mathcal{G}_t$ and the edge manipulation example can be:

$$\mathcal{G}'_{t+1} = \begin{cases} (\mathcal{V}_t, \mathcal{E}_t), & \mathcal{L}_{\mathcal{G}'_t} - \mathcal{L}_{\mathcal{G}_{t,c_i}} \leq 0 \\ (\mathcal{V}_t, \mathcal{E}_t \cup c_i), & c_i \notin \mathcal{E}_t, \mathcal{L}_{\mathcal{G}'_t} - \mathcal{L}_{\mathcal{G}_{t,c_i}} > 0 \\ (\mathcal{V}_t, \mathcal{E}_t \setminus c_i), & \mathcal{L}_{\mathcal{G}'_t} - \mathcal{L}_{\mathcal{G}_{t,c_i}} > 0, \end{cases}$$

where $\mathcal{L}_{\mathcal{G}'_t}$ is the objective value, $\mathcal{L}_{\mathcal{G}'_{t,c_i}}$ is the objective value of the updated graph (add/remove candidate edge c_i). Repeat the process until it meets $\mathcal{D}(\mathcal{G}, \mathcal{G}') > \Delta$, or already traversed all the candidate edges.

To predict the target opinion under adversarial perturbation (feature or structure), we need to solve the generalized problem formulate as:

$$\min_{\theta} \left\{ \max_{\mathcal{D}(\cdot)\leq\{\gamma \text{ or } \Delta\}} \mathcal{L}(\cdot) \right\} \tag{12}$$

where $\mathcal{L}(\cdot)$ indicates the final objective and $\mathcal{D}$ is a certain distance metric on graph perturbation. An inner maximization constructing an adversarial perturbed graph, and outer minimization is solved by our proposed robust inference model. Now we discuss how to formulate the objective in detail.

C. Adv-COI Inference Algorithm

1) Formulation of Unknown Opinion Inference: In the adversarial attack, we see the opinions of variables in set S as the **target opinions** whose opinions are unknown, and in $\overline{S}$ as the **attacker opinions** whose opinions might be perturbed. Without the constraints based on the logic rules, the joint PDF of all the variables has the following form:

$$\begin{aligned} &\text{Prob}(\mathbf{p}, \mathbf{b}_{\overline{S}}, \mathbf{X}_S, \mathbf{X}_{\overline{S}}; \boldsymbol{\omega}_S, \boldsymbol{\omega}_{\overline{S}}, \boldsymbol{\omega}_0, p_0) = \\ &\prod_{i=1}^{M} \Big\{ \Big(\text{Beta}(p_{\overline{S}_i}; \omega_{\overline{S}_i})\Big)^{1-b_{\overline{S}_i}} \Big(\text{Beta}(p_{\overline{S}_i}; \omega_0)\Big)^{b_{\overline{S}_i}} \cdot \text{Bin}(X_{\overline{S}_i}; p_{\overline{S}_i}) \\ &\text{Bin}(b_{\overline{S}_i}; p_0) \Big\} \prod_{l=1}^{N} \text{Bin}(X_{S_l}; p_{S_l}) \text{Beta}(p_{S_l}; \omega_{S_l}), \end{aligned} \tag{13}$$

where $\text{Bin}(\cdot)$ and $\text{Beta}(\cdot)$ refer to PDF of a Binomial distribution and a Beta distribution, respectively. For simplicity, we denote the parameters as $\theta = \{\boldsymbol{\omega}_{\overline{S}}, \boldsymbol{\omega}_0, p_0\}$. In the above PDF, if we do not consider logical relationships, the *input variables* and *output variables are independent*; thus we cannot predict the target opinions ω_S based on the input evidence. The goal is to identify the opinion vector $\boldsymbol{\omega}_S$ and the adversarial attacker indicator vector $b_{\overline{S}}$, such that the likelihood $\text{Prob}(\mathbf{X}_{\overline{S}}; \boldsymbol{\omega}_S, \theta)$ is maximized, subject to the constraints defined by the set of logic rules $\mathcal{R}$. We therefore integrate the logic rules $\mathcal{R}$ to model the dependency between latent probability variables p_{S_i} and $p_{\overline{S}_i}$. We apply a commonly used strategy that imposes the rule constraints on $\text{Prob}(\mathbf{p}, \mathbf{b}_{\overline{S}} | \mathbf{X}_{\overline{S}}; \boldsymbol{\omega}_S, \theta)$ through an expectation operator. By the definition in Eq. (9), for each rule, r_k, we expect that the distance to satisfaction $d_k(\cdot)$ close to zero, $\mathbb{E}_{\text{Prob}(\mathbf{p}, \mathbf{b}_{\overline{S}} | \mathbf{X}_{\overline{S}}; \boldsymbol{\omega}_S, \theta)}[d_k(\mathbf{p}, \mathbf{b}_{\overline{S}})] = 0$, with a confidence measured by the weight ρ_k. The unknown opinion prediction problem can be formulated as a maximization problem based on a log constrained likelihood, $\mathbb{L}(\boldsymbol{\omega}_S)$ by:

$$\begin{aligned} \max_{\boldsymbol{\omega}_S} \mathbb{L}(\boldsymbol{\omega}_S) &= \max_{\boldsymbol{\omega}_S} \log \text{Prob}(\mathbf{X}_{\overline{S}}; \boldsymbol{\omega}_S, \theta) \\ s.t.\ \mathbb{E}_{\text{Prob}(\mathbf{p}, \mathbf{b}_{\overline{S}} | \mathbf{X}_{\overline{S}}; \boldsymbol{\omega}_S, \theta)} &\Big[\rho_k \cdot d_k(\mathbf{p}, \mathbf{b}_{\overline{S}}) \Big] \leq \xi_k, \\ &\|\boldsymbol{\xi}\| \leq \epsilon, k = 1, \cdots, K, \end{aligned} \tag{14}$$

where ξ_k is a vector of slack variables. We allow small violations with slack variables ξ_k on the logic rules whose norm is bounded by $\epsilon \geq 0$.

2) Approximate Expectation Estimation: The main unknown opinion prediction under adversarial attack problem in Eq. (14) has two challenging computational complexity issues: i) the integral term $\mathbb{E}_{\text{Prob}(\mathbf{p}, \mathbf{b}_{\overline{S}} | \mathbf{X}_{\overline{S}}; \boldsymbol{\omega}_S, \theta)}[d_k(\mathbf{p}, \mathbf{b}_{\overline{S}})]$ is analytically intractable; and ii) the dimension of target opinion $\boldsymbol{\omega}_S$ is often large in scale (as large as the number of edges/nodes of our real world datasets). To solve the maximization problem in Eq. (14) and find an analytically tractable solution, we adopt ***posterior regularization (PR)*** [4], a probabilistic framework for structural relational learning. Via applying PR, Adv-COI learns a simpler density function $q(\mathbf{p}, \mathbf{b}_{\overline{S}})$ that fits the rules while staying close to the posterior PDF ($\text{Prob}(\mathbf{p}, \mathbf{b}_{\overline{S}} | \mathbf{X}_{\overline{S}}; \boldsymbol{\omega}_S, \theta)$).

We propose an efficient approximate expectation estimation algorithm reducing computational complexity. Due to the space constraint, we directly show the analytic form of $q(\mathbf{p}, \mathbf{b}_{\overline{S}})$ and the element-wise solution of $\omega_S = (\alpha_S, \beta_S)$:

$$q(\mathbf{p}, \mathbf{b}_{\overline{S}}) \propto \text{Prob}(\mathbf{p}, \mathbf{b}_{\overline{S}} | \mathbf{X}_{\overline{S}}; \boldsymbol{\omega}_S, \theta) \cdot \exp\left\{ -\textstyle\sum_{k=1}^{K} \rho_k d_k(\mathbf{p}, \mathbf{b}_{\overline{S}}) \right\}. \tag{15}$$

$$\max_{\alpha_{S_i}>0, \beta_{S_i}>0} \mathbb{E}_{q^{l+1}} \Big[\log \text{Beta}(p_{S_i} | \alpha_{S_i}, \beta_{S_i}) \Big] + const. \tag{16}$$

Opinion Inference Algorithm: Now we present an efficient approximate expectation estimation algorithm with less computational complexity of $\{\mathbb{E}_{q^{l+1}}[\log p_{S_i}],\ \mathbb{E}_{q^{l+1}}[\log(1-p_{S_i})] \mid i=1,\cdots,N\}$. Because the computation of these expectation terms is analytically intractable, we adopt a common approximation approach: $\mathbf{p}^*$ and $\mathbf{b}^*_{\overline{S}}$ represent the values at the most probable setting of $\mathbf{p}$ and $\mathbf{b}_{\overline{S}}$ with the current opinion $\boldsymbol{\omega}_S$. The expectations terms $\mathbb{E}_{q^{l+1}}[\log p_{S_i}]$ and $\mathbb{E}_{q^{l+1}}[\log(1-p_{S_i})$ can be approximated as $\log \mathbf{p}^*_{S_i}$ and $\log(1-\mathbf{p}^*_{S_i})$, respectively. The most probable values to predict $\mathbf{p}^*$ and $\mathbf{b}^*_{\overline{S}}$ can be estimated by solving the following optimization problem (from the analytical solution in Eq. (15)):

$$\begin{aligned}\mathbf{p}^*, \mathbf{b}^*_{\overline{S}} &= \arg\min_{\mathbf{p},\mathbf{b}_{\overline{S}}} -\log q(\mathbf{p}, \mathbf{b}_{\overline{S}}) \qquad (17)\\ &= \arg\min_{\mathbf{p},\mathbf{b}_{\overline{S}}} -\log \text{Prob}(\mathbf{p}, \mathbf{b}_{\overline{S}}|\mathbf{X}_{\overline{S}}) + \sum_{k=1}^{K}\rho_k d_k(\mathbf{p}, \mathbf{b}_{\overline{S}}),\end{aligned}$$

where parameters $\boldsymbol{\omega}_S$ and θ are omitted in the Prob$(\cdot)$ function for simplicity. To achieve high scalability, we propose an efficient algorithm using the alternating direction method of multipliers (ADMM) [9] to solve the above problem. The ADMM has three main steps: i) ***forming and initializing*** local copies of the variables in each logic rule by constraining the local copies to be equal to the global variables; ii) ***decomposing*** the problem into independent subproblems; and iii) ***block-wise updating*** until converging to a consensus on the optimum.

Let $\hat{\mathbf{p}}_k$ and $\hat{\mathbf{b}}_{\overline{S},k}$ be local copies of the global variables $\mathbf{p}$ and $\mathbf{b}_{\overline{S}}$ that are involved in the logic rule r_k. $\mathbf{p}_k$ and $\mathbf{b}_{\overline{S},k}$ be the variables in $\mathbf{p}$ and $\mathbf{b}_{\overline{S}}$ that correspond to $\hat{\mathbf{p}}_k$ and $\hat{\mathbf{b}}_{\overline{S},k}$, $(k=1,\cdots K)$, respectively. Finally out main problem based on the ADMM is formulated as follows:

$$\begin{aligned}&\min_{\hat{\mathbf{p}},\hat{\mathbf{b}}_{\overline{S}},\mathbf{p},\mathbf{b}_{\mathbf{y}}} \Big\{ -\log \text{Prob}(\mathbf{p}, \mathbf{b}_{\mathbf{y}}|\mathbf{X}_{\overline{S}}) + \sum_{k=1}^{K}\rho_k d_k(\mathbf{p}, \mathbf{b}_{\overline{S}})\Big\},\\ &s.t.\ \hat{\mathbf{p}}_k = \mathbf{p}_k, \hat{\mathbf{b}}_{\overline{S},k} = \mathbf{b}_{\overline{S},k}\ \forall k = 1,\cdots,K.\end{aligned}$$

The augmented Lagrangian with penalty κ and Lagrange multipliers λ and θ of the above objective function is:

$$\begin{aligned}&\mathcal{L}(\hat{\mathbf{p}}, \mathbf{p}, \hat{\mathbf{b}}_{\overline{S}}, \mathbf{b}_{\overline{S}}, \boldsymbol{\lambda}, \boldsymbol{\eta}_{\overline{S}}) \qquad (18)\\ =\ & -\log \text{Prob}(\mathbf{p}, \mathbf{b}_{\overline{S}}|\mathbf{X}_{\overline{S}}) + \sum_{k=1}^{K}\Big[\rho_k d_k(\mathbf{p}, \mathbf{b}_{\overline{S}}) +\\ +\ & \frac{1}{2\kappa}\|\hat{\mathbf{p}}_k - \mathbf{p}_k + \kappa\boldsymbol{\lambda}_k\|_2^2 + \frac{1}{2\kappa}\|\hat{\mathbf{b}}_{\overline{S},k} - \mathbf{b}_{\overline{S},k} + \kappa\boldsymbol{\eta}_{\overline{S},k}\|_2^2\Big],\end{aligned}$$

where $\kappa > 0$. ADMM finds a saddle point of the Lagrangian $\mathcal{L}(\hat{\mathbf{p}}, \hat{\mathbf{b}}_{\overline{S}}, \boldsymbol{\lambda}, \boldsymbol{\eta}_{\overline{S}}, \mathbf{p}, \mathbf{b}_{\overline{S}})$ by updating the four blocks of variables at each iteration t:

$$\boldsymbol{\lambda}^t_k = \boldsymbol{\lambda}^{t-1}_k + \frac{1}{\kappa}(\hat{\mathbf{p}}^{t-1}_k - \mathbf{p}^{t-1}_k) \qquad (19)$$

$$\boldsymbol{\eta}^t_{S,k} = \boldsymbol{\eta}^{t-1}_{S,k} + \frac{1}{\kappa}(\hat{\mathbf{b}}^{t-1}_{\overline{S},k} - \mathbf{b}^{t-1}_{\overline{S},k}) \qquad (20)$$

$$\begin{aligned}\hat{\mathbf{p}}^t_k, \hat{\mathbf{b}}^t_{\overline{S},k} &= \arg\min_{\hat{\mathbf{p}}_k,\hat{\mathbf{b}}_{\overline{\mathbf{S}},\mathbf{k}}} \rho_k d_k(\mathbf{p}^{t-1}, \mathbf{b}^t_{\overline{S}}) + \frac{1}{2\kappa}\|\hat{\mathbf{p}}_k - \mathbf{p}^{t-1}_k + \kappa\boldsymbol{\lambda}^t_k\|_2^2\\ &+ \frac{1}{2\kappa}\|\hat{\mathbf{b}}_{\overline{S},k} - \mathbf{b}^{t-1}_{\overline{S},k} + \kappa\boldsymbol{\eta}^t_{\overline{S},k}\|_2^2,\ \forall k = 1,\cdots,K \qquad (21)\\ \mathbf{p}^t, \mathbf{b}^t_{\overline{S}} &= \arg\min_{\mathbf{p},\mathbf{b}_{\overline{S}}} \mathcal{L}(\hat{\mathbf{p}}^t, \mathbf{p}, \hat{\mathbf{b}}^t_{\overline{S}}, \mathbf{b}_{\overline{S}}, \boldsymbol{\lambda}^t, \boldsymbol{\eta}^t_{\overline{S}}). \qquad (22)\end{aligned}$$

The updates of the ADMM make sure that $\mathbf{p}$ and $\mathbf{b}_{\overline{S}}$ converge to the global optimums $\mathbf{p}^*$ and $\mathbf{b}^*_{\overline{S}}$, assuming that there exist feasible assignments to $\mathbf{p}$ and $\mathbf{b}_S$. The global variables related problem in Eq. (22) has an analytical solution ensuring that the gradient of the objective function is 0. We can efficiently solve the local variables related problem in Eq. (21) via adopting the algorithm designed in [3].

Generally the posterior of the inference method is not trained. We propose an iterative inference algorithm Adv-COI for uncertainty-based opinion prediction. Adv-COI (i.e., a "black-box" inference method) does not use the ground truth of testing opinions (i.e., target opinions) during the construction of adversarial opinions and structure perturbation, or infer of the unknown opinions. As we defined in Section IV-A, $\mathbf{p} = \Gamma(\mathbf{p}_S, \mathbf{p}_{\overline{S}})$, and the opinions of variables in $\overline{S}$ is given, and the variables in S are unknown. During the inference, the Adv-COI learns a soft indicator $\mathbf{b}_{\overline{S}}$ for each training opinions. Via learning such a vector $\mathbf{b}_{\overline{S}}$, the Adv-COI makes sure to decrease the influence of perturbed adversarial opinions. Our experimental results also prove that the proposed defense mechanism is effective on both clean and perturbed data.

Following the recipe of adversarial defense, we formulate the opinion inference problem as:

$$\min_{\mathbf{p},\mathbf{b}_{\overline{S}}} \Big\{ \max_{\mathcal{D}(\cdot,\cdot)\le\gamma} \mathcal{L}(\hat{\mathbf{p}}, \mathbf{p}, \hat{\mathbf{b}}_{\overline{S}}, \mathbf{b}_{\overline{S}}, \boldsymbol{\lambda}, \boldsymbol{\eta}_{\overline{S}})\Big\} \qquad (23)$$

where $\mathcal{D}(\cdot,\cdot)$ indicates a distance metric of feature (opinion) or a structural adversarial perturbation.

Algorithm 1: Adv-COI

Input: $\boldsymbol{\omega}_{\overline{S}}, \mathbf{X}_{\overline{S}}, \mathcal{R}, \omega_0, p_0, \alpha, \gamma, \mathcal{G}, k$
Output: $\boldsymbol{\omega}_S$

1 *#1. Generate adversarial opinions and/or perturbed factor graph.*
2 $\mathcal{G}', p^{\mathbf{adv}}_{\overline{S}}$:= **gen_adv_samples**$(\boldsymbol{\omega}_{\overline{S}}, \mathbf{X}_{\overline{S}}, \mathcal{R}, \omega_0, p_0, \alpha, \gamma, \mathcal{G}, k)$;
3 *#2. Adversarial collective opinion inference.*
4 $\boldsymbol{\omega}_S$ = **collective_opinion_infer**$(p^{\mathbf{adv}}_{\overline{S}}, \boldsymbol{\omega}_{\overline{S}}, \mathbf{X}_{\overline{S}}, \mathcal{R}, \omega_0, p_0, \mathcal{G}')$;
5 **return** $\boldsymbol{\omega}_S$

3) Adv-COI Adversarial Attacks: In Eq. (23), the **inner maximization** problem is equivalent to constructing a stronger adversarial sample.

In fact, solving Eq. (23) to a saddle point can be done by performing multiple PGD steps or structural perturbation methods introduced in Section IV-B. Algorithm 1 shows the pseudo code of the Adv-COI with two main steps: (1) **Line 2** generates adversarial opinions via perturb opinions and/or factor graph structure; and (2) **Line 4** infers target opinion under adversarial attack via Algorithm 2.

ℓ_∞-PGD adversarial attack: We want to construct stronger adversarial opinions, by distorting the truth probability $p^{adv}_{\overline{S}}$ of the training opinions, which maximize the total weighted distance to satisfaction of the rules related to the target (test) opinions. In **Line 2**, Algorithm 1 generates training (attacker)

opinions by conducting PGD attacks in γ-ball distortion, where we directly control the perturbation size with γ:

$$p^{adv}_{\overline{S}} = \{p_{\overline{S}} + \delta^* | \delta^* = \arg\max_{\|\delta\|_\infty \leq \gamma} \mathcal{L}(\hat{\mathbf{p}}, \mathbf{p}, \hat{\mathbf{b}}_{\overline{S}}, \mathbf{b}_{\overline{S}}, \boldsymbol{\lambda}, \boldsymbol{\eta}_{\overline{S}})\}. \quad (24)$$

where $\delta = p^{adv}_{\overline{S}} - p_{\overline{S}}$. Starting from $p^0_{\overline{S}}$, PGD attack conducts projected gradient decent iteratively to update the following adversarial example:

$$p^{t+1}_{\overline{S}} = \mathbf{Clip}_{p_{\overline{S}},\gamma}\{p^t_{\overline{S}} + \alpha \cdot \mathbf{sign}(\nabla_{p_{\overline{S}}} \mathcal{L}(\hat{\mathbf{p}}^t, \mathbf{p}^t, \hat{\mathbf{b}}^t_{\overline{S}}, \mathbf{b}^t_{\overline{S}}, \boldsymbol{\lambda}^t, \eta^t_{\overline{S}}))\}, \quad (25)$$

where $\mathbf{Clip}_{p_{\overline{S}},\gamma}(\cdot)$ element-wise clips the input into the range $[p_{\overline{S}} - \gamma, p_{\overline{S}} + \gamma]$, such that $p_{\overline{S}} - \gamma, p_{\overline{S}} + \gamma \in [0, 1]$, "sign" essentially enforcing the max norm constraint. In Eq. (22), when other other variables are fixed, it is tractable to calculate $\nabla_{p_{\overline{S}}} \mathcal{L}(\cdot)$. $p_{\overline{S}}$ indicates the clean data, $\alpha = 0.02$ indicates each value changes with 0.02 unit during each update.

Structure adversarial attack: We can generate a perturbed graph applying the greedy algorithm proposed in Section IV-B to solve the problem in Eq. (26).

$$\max_{\mathcal{G}', \mathcal{D}(\mathcal{G}, \mathcal{G}') \leq \Delta} \mathcal{L}_{\mathcal{G}'}(\hat{\mathbf{p}}, \mathbf{p}, \hat{\mathbf{b}}_{\overline{S}}, \mathbf{b}_{\overline{S}}, \boldsymbol{\lambda}, \boldsymbol{\eta}_{\overline{S}}), \quad (26)$$

where $\mathcal{L}_{\mathcal{G}'}$ represents the associated perturbed factor graph, Δ is the maximum unnoticeable change to the factor graph.

Algorithm 2: Adversarial Defense Inference on Opinion Prediction

Input: $p^{\mathbf{adv}}_{\overline{S}}, \boldsymbol{\omega}_{\overline{S}}, \mathbf{X}_{\overline{S}}, \mathcal{R}, \omega_0, \alpha, \mathcal{G}'$
Output: $\boldsymbol{\omega}_S$

1 $l = 1$;
2 Initialize $\boldsymbol{\omega}^l_S$;
3 **repeat**
4 Update $q^l(\mathbf{p}, \mathbf{b}_{\overline{S}})$ via Eq. (15),where $\mathbf{p} = \Gamma(\mathbf{p}_S, \mathbf{p}^{\mathrm{adv}}_{\overline{S}})$
5 $t = 1$;
6 Initialize $\mathbf{p}_S, \mathbf{b}_{\overline{S}}$;
7 Initialize $\hat{\mathbf{p}}^t_k$ and $\hat{\mathbf{b}}^t_{\overline{S},k}$ as copies of the probability variables, $\mathbf{p}^t_k$,$\mathbf{p}^{\mathbf{adv}}_{\overline{S}}$ and $\mathbf{b}^t_{\overline{S},k}$, that appear in the k-th rule in $\mathcal{R}$
8 Initialize Lagrange multipliers $\boldsymbol{\lambda}_k$ and $\boldsymbol{\eta}_{\overline{S},k}$,$k = 1, \cdots, K$;
9 **repeat**
10 $t = t + 1$
11 Update $\boldsymbol{\lambda}^t_k$ via Eq. (19), $k = 1, \cdots, K$;
12 Update $\boldsymbol{\eta}^t_{\overline{S}}$ via Eq. (20), $k = 1, \cdots, K$;
13 Update $\hat{\mathbf{p}}^t_k$ and $\hat{\mathbf{b}}^t_{\overline{S},k}$ by solving the problem in Eq. (21), $k = 1, \cdots, K$;
14 Update $\mathbf{p}^t, \mathbf{b}^t_{\overline{S}}$ by solving the problem in Eq. (22);
15 **until** *convergence*
16 $l = l + 1$;
17 **for** $i = 1, \cdots, N$ **do**
18 Update $\omega^l_{S_i}$ by solving Eq.(16);
19 **until** *convergence*
20 **return** $\boldsymbol{\omega}^l_S$

4) Adv-COI Opinion Prediction with Defense: Adv-COI is a robust probabilistic inference method against adversarial attacks, and simultaneously infers unknown opinions and performs the adversarial learning. The key steps of opinion inference algorithm are summarized in Algorithm 2.

The outer loop relates to the modified Expectation Maximization (EM) procedure. The E′−Step is implemented by Line 4. The M−Step is implemented from Lines 5 to 18. In particular, Lines 5 to 15 implement the ADMM procedure to estimate the most probable values $\mathbf{p}^*$ by solving the optimization problem in Eq. (17). The estimated probable values $\mathbf{p}^*$ are used to approximate $\mathbb{E}_{q^{l+1}}[\log p_{S_i}]$ and $\mathbb{E}_{q^{l+1}}[\log(1 - p_{S_i})$ as $\log \hat{p}_{S_i}$ and $\log(1 - \hat{p}_{S_i})$, respectively, which are then used to implement the M−Step in Lines 17 and 18.

Complexity Analysis: The time complexity of Algorithm 2 is dominated by Lines 13, 14, and 18. Line 18 needs to solve the optimization problem in Eq. (16) which is the same as the Maximum Likelihood Estimation (MLE) problem of a Beta distribution and can be solved using the method of moments [2] with $O(1)$. Line 13 needs to solve K problems in Eq. (21) that can be solved using the algorithm designed in [3] with $O(KP)$, where P is the maximum number of variables involved in the logic rules, $\mathcal{R}$. Line 14 needs to solve the optimization problem in Eq. (22) whose analytical solution can be obtained in $O(N + 2M)$. Let L_1 and L_2 be the numbers of iterations on the outer and inner loops, respectively. The overall complexity of Algorithm 2 is $O(L_1 \cdot L_2 \cdot (K + N + 2M + KP))$. L_1, L_2 are small numbers and K problems in Line 13 can be computed in parallel. If we have sufficient cores with count C such that $O(K/C) \approx O(1)$, we then have $O(L_1 \cdot L_2 \cdot (K + N + 2M + P))$, which is linear with respect to N, M, and K. This proves scalability of our proposed algorithm with large-scale network data.

TABLE II: Dataset statistics

Application	Dataset	# Nodes	# Edges	Avg. Degree
Congestion Prediction	Philadelphia (PA)	603	708	1.17
	Washington D.C. (DC)	1,383	1,878	1.35
Trust Inference	Epinions (EP)	5,000	9,288	1.85
Sybils User Detection	Facebook (FB)	8,078	372,936	23.08
	Enron (EN)	67,392	743,244	5.51
	Slashdot (SD)	164,336	2,018,920	6.14

V. Experimental Setup

We evaluate quality, scalability, and real-world utility of the Adv-COI with other strong baselines based on six semi-synthetic and real world datasets. All experiments are tested on 56 CPUs of Intel Xeon (R) E5-2680 with 251G of RAM.

A. Datasets

We validate the Adv-COI on four semi-synthetic datasets and two real world datasets representing different real world applications. Dataset statistics are summarized in Table II.

1) **DC. and PA. road traffic datasets:** These datasets are the collected traffic data from June 1, 2013 to March 31, 2014 across two cities from INRIX [1], Washington D.C. and Philadelphia (PA), as summarized in Table II. The raw INRIX dataset provides traffic speed and reference speed information for each road link per hour interval. A reference speed is defined as the "uncongested free flow speed" for each road segment [1]. It is calculated based upon the 60-th percentile of the measured speed for all time periods over a few years, where the reference speed serves as a threshold separating two traffic states, *congested* vs. *uncongested*. The road traffic dataset for each of the two cities has 43 weeks in total. An hour is represented by a specific combination of hour of day ($h \in \{8, \cdots, 21\}$), day of week ($d \in \{1, \cdots, 5\}$), and week ($w \in \{1, \cdots, 43\}$): (h, d, w).

Estimation of opinions of the training and testing edges. For each road traffic dataset, the opinion of a specific (training or testing) link i at an hour (h, d, w) is estimated based

[1] http://inrix.com/publicsector.asp

on the observations of the same hour in previous T weeks $\{x_{i,h,d,w}, x_{i,h,d,w-1}, \cdots, x_{i,h,d,w-T+1}\}$ as the evidence, where $x_{i,h,d,w}$ refers to the congestion observation (0 or 1) of the link i at hour (h, d, w) and T refers to a predefined time window size. The belief, disbelief, and uncertainty mass variables b_{x_i}, d_{x_i}, and u_{x_i} of a specific link i are estimated as:

$$b_{x_i} = \frac{\sum_{t=0}^{T-1} x_{i,h,d,w-t}}{T+W}, \quad u_{x_i} = \frac{W}{T+W}$$
$$d_{x_i} = \frac{T - \sum_{t=0}^{T-1} x_{i,h,d,w-t}}{T+W} \tag{27}$$

where we set the non-informative prior weight (i.e., an amount of uncertain evidence) $W = 2$ and the base rate (i.e., prior knowledge) $a = 0.5$. For the other datasets, the opinion of each training/testing edge/nodes is estimated based on the T observations similar to the above.

2) **Epinions:** This dataset[2] represents a who-trust-who in an online social network. Epinion trust directed network has 5,000 users (i.e., nodes) and 9,288 trust relationships (i.e., edges). As there are no ground truth opinions available from the dataset, we use a benchmark simulation model [8], [12] to generate synthetic opinions.

Performance evaluation: After conducting T realizations, each edge then has up to T trust observations and its opinion can be estimated based on its trust observations. We consider a set of candidate values of $T \in \{8, 9, 10, 11\}$, corresponding to different uncertainty ranges that will be explained in the traffic dataset part.

3) **Social Networks with Synthesized Sybils Attack:** We utilize three social networks used in [10], [11], for example, Facebook, Enron, and Slashdot representing different application scenarios. We obtained these datasets from SNAP[3]. A node in Facebook dataset represents a user on Facebook, and two nodes are connected if they are friends. A node in Enron dataset represents an email address, and an edge between two nodes indicates at least one email was exchanged between two corresponding email addresses. Slashdot is a technology-related news website, which allows users to tag each other as friends or foes. Slashdot network thus contains friend/foe links between users. We follow the method of synthesizing the Sybil attack in different scenarios [10], [11]. We use a real social graph as the Benign region while synthesizing the Sybil region (for each social network, we use it as the Benign region and replicate it as a Sybil region) and add attack edges between the two regions uniformly at random. We label the observation of the nodes in the Sybil region to "1" at time stamp t=1, "0" to the nodes in the Benign region. In the exploration step, we duplicate the observations of each node and process T realizations, and then we randomly swap observations of 1% of nodes each realization. Set $T = 10$ in these three datasets. We also try different numbers attacking edges between the Benign region and the Sybil region, {1000, 5000, 10000, 15000, 20000} which makes our prediction task more challenging.

[2]http://www.trustlet.org/downloaded_epinions.html

[3]http://snap.stanford.edu/data/index.html

TABLE III: Logic rules used in the experiments

Epinions Rules [13]	
$\text{TRUSTS}(A,B) \wedge \text{TRUSTS}(B,C) \rightarrow \text{TRUSTS}(A,C)$ $\neg\text{TRUSTS}(A,B) \wedge \text{TRUSTS}(B,C) \rightarrow \neg\text{TRUSTS}(A,C)$ $\text{TRUSTS}(A,B) \wedge \neg\text{TRUSTS}(B,C) \rightarrow \neg\text{TRUSTS}(A,C)$	A,B and C are users, Trust$(\cdot,\cdot)$ indicates their trust relationship.
Road Traffic Rules [14]	
$\text{NEIGHBOR}(E_1,E_2) \wedge \text{CONGESTED}(E_1) \rightarrow \text{CONGESTED}(E_2)$	If E_1 is a congested road section and E_2 is its upper neighbor, then E_2 is likely congested.
Sybils Attack Rules [10]	
$\text{LINKED}(User_1, User_2) \wedge \text{HASLABEL}(User_1, Type_A)$ $\rightarrow \text{HASLABEL}(User_2, Type_A)$	Two linked network entities share the same label with a high probability.

For all these datasets, the testing opinions on the edges/nodes are randomly selected with the percentages or test ratios $(TR) := \frac{N}{N+M} \times 100\% \in \{10\%, 20\%, 30\%, 40\%, 50\%\}$ and are predicted based on the observations and known opinions of the other edges/nodes which are the training edges/nodes.

B. Experimental Setup

1) ***Baselines:*** In essence, our method is inspired by SL [6], PSL [3] (Section III), and CSL [12], so these three methods are natural baselines. We also compare Adv-COI with GCN-VAE-opinion [4] (or GCN-VAE) [15] is a state-of-the-art DL-based method, and a naive Baseline0, which predicts the opinion of each target testing edge (node) as randomly (1,0,0) or (0,1,0), always true or false with zero uncertainty.

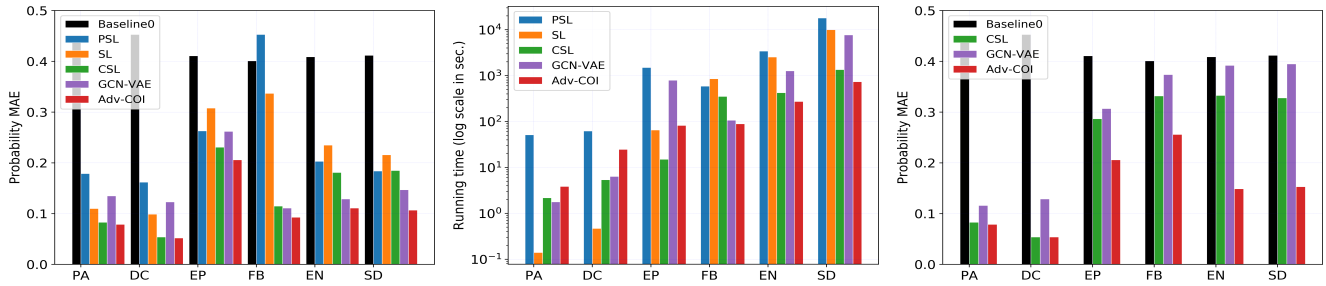
2) ***Parameter Settings:*** SL only has one hyperparameter that is the maximum length of its independent paths, and we try different settings $[3, \cdots, 20]$ and for each dataset we keep the settings that return the best result. Adv-COI, PSL and CSL require logical rules as additional inputs for reasoning. We consider logic rules from the related papers corresponding to the datasets. Table III lists the logic rules of each dataset. We set all rule weights to 1.0,equally important, and $\omega_0 = (1, 1)$ and $p_0 = 0.5$. GCN-AVE-opinion we use the recommended settings from the paper: $\lambda = 0.01$ (trade-off parameter), $\eta = 0.001$ (learning rate), and $P = 16$ (dimensionality of latent encoded vectors), and dropout rate=0.1.

3) ***Performance Metrics:*** **(1) Expected truth probability Mean Absolute Error (MAE).** Based on Eq. (27), the uncertainty mass, u_{S_i}, for each training or testing opinion is a known and constant value after the window size T is predefined, without the actual observations of this link. For this reason, the empirical analysis is focused on the comparison between Adv-COI and baselines in terms of *Expected truth probability MAE* (denoted as Probability MAE or EP-MAE). EP-MAE is defined as:

$$\text{EP-MAE}(\boldsymbol{\omega_S}) = \frac{1}{N} \sum\nolimits_{i=1}^{N} \left| \frac{\alpha_{S_i}}{\alpha_{S_i} + \beta_{S_i}} - \frac{\alpha^{\star}_{S_i}}{\alpha^{\star}_{S_i} + \beta^{\star}_{S_i}} \right| \tag{28}$$

where $\omega_{S_i} = (\alpha_{S_i}, \beta_{S_i})$ and $\omega^{\star}_{S_i} = (\alpha^{\star}_{S_i}, \beta^{\star}_{S_i})$ refers to the predicted and true opinions of a target variable S_i, respectively, and $\frac{\alpha_{S_i}}{\alpha_{S_i}+\beta_{S_i}}$ refers to the expected truth probability (or the expected belief) of the opinion ω_{S_i}. Expected probability MAE is calculated as the average absolute difference between the

[4]refer to: https://github.com/zxj32/GCN-VAE-opinion



(a) Expected truth probability MAE for clean data (b) Average running time (c) Expected truth probability MAE for perturbed data with $\gamma = 0.01$

Fig. 2: Expected truth probability MAE and average running time on the clean and perturbed data

estimated expected belief and the true expected belief on all testing opinions. **(2) Adversarial sample transferability.** We measure the adversarial sample correlation between different models where the correlation is defined by:

$$\rho_{A\mapsto B} = \frac{\textbf{Mae}[B|A] - \textbf{Mae}[B]}{\textbf{Mae}[B|B] - \textbf{Mae}[B]} \tag{29}$$

where $\rho_{A\mapsto B}$ measures the failing rate using source model A and target model B, **Mae**$[B]$ denotes the **MAE** of model B without attack, **Mae**$[B|A$ **(or** $B)]$ means the **MAE** under adversarial samples generated by model A (or B). Obviously, it is always easier to find adversarial examples through the target model itself, so we have **Mae**$[B|B] \geq$ **Mae**$[B|A]$ and thus $0 \leq \rho_{A\mapsto B} \leq 1$. However, $\rho_{A\mapsto B} = \rho_{B\mapsto A}$ is not necessarily true, so the correlation matrix is not likely to be symmetric. **(3) Average running time.** We compare the running time of the Adv-COI and other baseline methods on the real world dataset experiments. **(4) Prediction accuracy:** measures node/edge level prediction accuracy where the decisions are made based on the predicted opinions on real world dataset experiments.

VI. Experiments Results

A. Uncertainty-based Opinion Prediction

For validating the performance of the Adv-COI, we conducted comparative performance analysis by comparing it with the baselines or competitive counterparts in terms of the metrics based on the six real world datasets in Section V-B.

1) ***Opinion Inference Performance****:* Fig. 2 (a) shows the expected truth probability MAE of the Adv-COI and baselines on the six datasets for clean data, where $T = 10$ for Epinions and three Sybil datasets, $T = 43$ for two traffic datasets. For Sybil datasets, the attacking edges between the Benign and Sybil region are 10,000.

From Fig. 2 (a), we can observe that the Adv-COI outperforms other five baseline methods on all the datasets with the minimum MAE generated. The overall performance order of the compared methods is **Adv-COI>GCN-VAE~CSL>SL >PSL**. Notice that the probability MAE values are higher on the Epinion dataset for Adv-COI, CSL and GCN-VAE methods compared to other dataset experiments. Because the trust relationships of test users' opinions on Epinion dataset is inferred from 2-hop neighbors opinion, which makes the prediction of opinions more challenging; however, Adv-COI still outperforms the counterparts.

2) ***Scalability****:* While exhibiting the above good prediction quality Adv-COI also shows high scalability on the dataset. The scale varies from the smallest Philly traffic network with 603 nodes (users) and 708 edges to the biggest one Slashdot network with 164,336 nodes and 2,018,920 edges. The number of rule instances changes from 1000 to 2 million. We show the average running time of our experiments in Fig. 2 (b). We neglect the running time of Baseline0 method (0.5~1.5 secs) and adversarial example generation part. As we have discussed in Section IV-C4, Adv-COI is linear w.r.t. the number of testing (N) and training (M) variables and the number logic rules (K), respectively, and has a lower complexity. Due to handling the large adjacency and feature matrix, GCN-VAE is slower for large graphs. The running time of SL increases exponentially when the network size increases, and is not scalable for large networks. The running time of PSL is high when it infers a large number of rule instances. Adv-COI and CSL scale nearly linear with respect to the network size. This shows high scalability of the Adv-COI on large datasets.

B. Evaluating Models Under l_∞-PGD Attacks

We compare the accuracy under the white box ℓ_∞-PGD attack. We set the maximum ℓ_∞ distortion to $\gamma \in \{0.0, 0.01, 0.03, 0.05, 0.07, 0.09, 0.2, 0.3, 0.4\}$ and report the probability MAE. We generate adversarial samples of real world datasets for attacking models of Adv-COI, Baseline0, CSL and GCN-VAE via following steps: (1) **Adv-COI** based on the Eq. (24) to generate perturbed data; (2) **Baseline0** based on random perturbations by adding noise on $[-\gamma, \gamma]$ to the attacker (training) observations (opinions) where maximum perturbation $|\delta_i| \leq \gamma$, and the following formula for perturbed attacker opinions: $p_{\overline{S}}^{adv} = \textbf{Clip}_{[0,1]}\{p_{\overline{S}} + \delta \mid |\delta_i| \leq \gamma\}$; (3) **CSL** based on the objective function Eq. (33) in [12]; and (4) **GCN-VAE** trained based on clean data and generated adversarial samples. All of our experiments are performed similar to the poisoning attack fashion.

Fig. 3 shows the effectiveness of the test ratio and distortion level γ to Adv-COI performance. As expected, we can observe the overall trend as follows. While the test ratio increases (the number of attackers decreases), the probability MAE of the Adv-COI decreases. As distortion level γ increases, the probability MAE of the Adv-COI increases as well. When the test ratio is low, the number of perturbed opinions is high. In this case, the adversarial attack is strong. So, when the test ratio is 10% and $\gamma = 0.4$, the probability MAE results in the

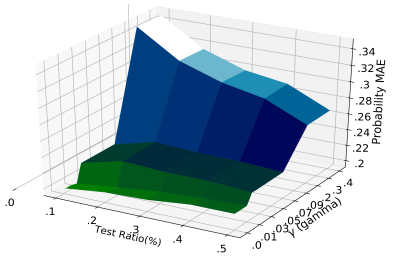
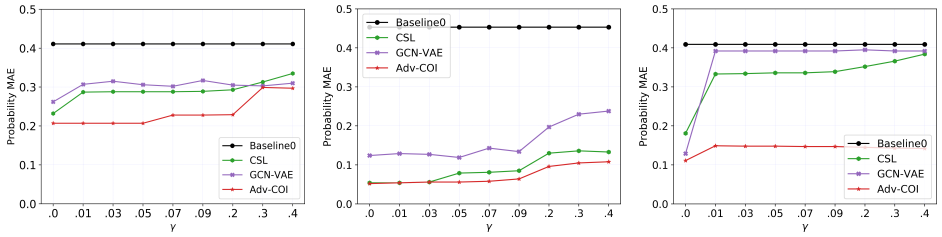


Fig. 3: Effectiveness of test ratio and gamma. The expected truth probability MAE of Adv-COI on Epinions dataset.



(a) Epinions (b) DC (c) Enron

Fig. 4: Probability MAE under ℓ_∞-PGD attack on three different datasets Epinions, DC and Enron (lower is better).

Adv-COI is the highest. When we fix the test ratio for all the datasets, the patterns of probability MAE are similar; so in our experiments, we only show the results of 30% test ratio.

Fig. 2 (c) shows the results of perturbed data with $\gamma = 0.01$. Compared to the results of clean data in Fig. 2 (a), for an unnoticeable perturbation with $\gamma = 0.01$, the probability MAE results of CSL and GCN-VAE increase dramatically for large datasets. In this perturbed data experiment, the Adv-COI outperforms baseline methods on all the dataset settings.

Fig. 4 shows the probability MAE results of the Adv-COI, Baseline0, CSL, and GCN-VAE methods under ℓ_∞-PGD attack on the Epinions, DC and Enron. Adv-COI outperforms CSL and GCN-VAE methods and shows strong robustness to ℓ_∞-PGD attacks.

Table IV shows the Sybil and Benign user detection (classification) accuracy of the Adv-COI and the best competitive baseline CSL under white box ℓ_∞-PGD attack with varying perturbation γ. We use the belief and uncertainty of predicted opinions in our decision making during the classification. Better accuracy is marked in bold. Notice that although our Adv-COI and CSL both incur accuracy drop, Adv-COI has high robust defense against adversarial attacks. Adv-COI leads to an improvement of 20~32% prediction accuracy compared to CSL under PGD attack with 0.01 distortion.

From Fig. 2, Fig. 4, and Table IV, we can observe that the Adv-COI has strong defense and high scalability to large network datasets, compared with those in baselines.

TABLE IV: Sybils attack prediction accuracy(%) under different $\gamma = \{0, 0.01, \cdots, 0.2\}$ levels of PGD attack.

Dataset	Defense	0	0.01	0.05	0.09	0.2
Facebook	Adv-COI	88.9	**60.8**	**60.1**	**62.2**	**63.3**
	CSL	**89.0**	50.6	50.6	50.6	50.5
Enron	Adv-COI	**86.7**	**82.1**	**82.2**	**82.3**	**82.5**
	CSL	75.3	50.2	50.2	50.1	50.1
Slashdot	Adv-COI	**87.3**	**81.6**	**81.8**	**81.8**	**82.4**
	CSL	73.1	49.5	49.5	49.4	49.4

C. Unnoticeable Structure Attack

After exploring how our designed feature attack method affects the different model, we will validate unnoticeable structure attacks on factor graph with different perturbation level $\Delta = \{0, 5, 10, \cdots, 90, 100\}$, indicating the number of perturbation on target nodes.

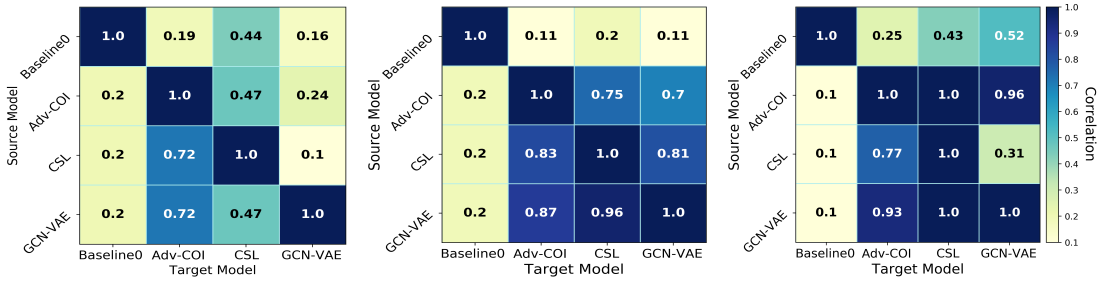
Fig. 6 (a) shows the probability MAE results of the target nodes in a direct structure attack with increasing perturbation level. Based on the proposed method in Section IV-B, from the testing nodes we randomly selected 20 nodes as target nodes, and conducted a direct structure perturbation on their neighbor nodes with different level perturbation. The Adv-COI shows strong resistance to structure attacks compared to CSL and GCN-VAE, especially with a high perturbation degree. Interestingly, when structure perturbation is small (< 20), the performance for all methods are stable because the Facebook network has a high average node degree.

We construct stronger adversarial data via conducting on both structure and opinion perturbation on target nodes, and the results are shown in Fig. 6 (b). Compared to the Fig. 6 (a), the performance of all models are dropping, demonstrating the combination of the structure and opinion attack are stronger effects to the models. However, Adv-COI still has resistant to the attack, and lags in an increase of MAE.

D. Black Box Transfer Attack Study

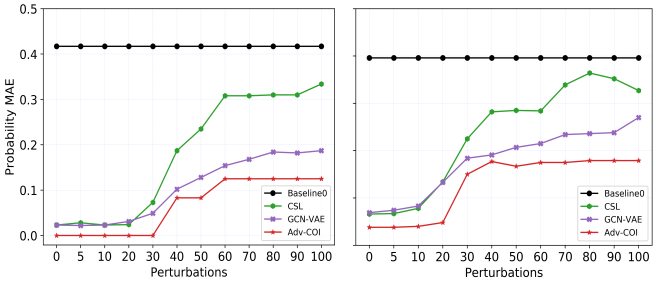
We measure the adversarial sample correlation between different models, namely Baseline0, Adv-COI, CSL and GCN-VAE. We employ the method called "transfer attack" [16]. We can imagine, the success rate of the transfer attack is directly linked with the similarity of the source/target models. We employ the same method in Section VI-B to generate adversarial samples for each model.

We select all combinations of source and target models, and calculate the correlation according to Eq. (29) on all datasets. Figs. 5 (a), (b) and (c) show black box transfer attack experimental results on the Epinions, DC and Enron, respectively. As expected, we can observe from Fig. 5 that Adv-COI and CSL are similar models, and GCN-VAE also has a weaker correlation with these two models. The above three methods are all robust to the same level of ℓ_∞ random noise attacks. From the correlation results, we can conclude that our ℓ_∞-PGD attack generates more accurate adversarial samples than ℓ_∞ random noise attack. Fig. 5 (a) shows Epinions results, $\rho_{\text{CSL}\mapsto\text{Adv-COI}}$ and $\rho_{\text{GCN-VAE}\mapsto\text{Adv-COI}}$ is around 0.7, indicating that the Adv-COI has a resistance to their black box attack to some extent. Figs. 5 (b) and (c) show the results based on DC and Enron. GCN-VAE and CSL have a correlation with Adv-COI. GCN-VAE has a weaker correlation with CSL, and has a strong correlation only in one direction.



(a) Epinions (b) DC (c) Enron

Fig. 5: Black box, transfer attack experiment results.



(a) Structure attack (b) Structure + Opinion attack

Fig. 6: Probability MAE of opinions on target nodes under different level of (a) structure and (b) structure-opinion (γ = 0.2) attack on Facebook dataset. (lower is better)

VII. Conclusion

We proposed the adversarial, collective opinion inference model, namely Adv-COI, which is a highly scalable, robust uncertainty-based opinion inference model. Adv-COI derives unknown opinions based on the known opinion probabilities, and to defend against adversarial attacks, on the opinion and graph structure, which have not been considered in the existing counterparts. The Adv-COI infers unknown opinions and learns adversarial defense simultaneously. We formulated the min-max problem in the Adv-COI to better learn the robust adversarial model as well as the unknown opinion derivation problem as an uncertainty minimization problem so that the Adv-COI can effectively predict unknown opinions with linear complexity. Our extensive experiments based on the six semi-synthetic and real world datasets demonstrated that the Adv-COI outperforms the state-of-the-art counterparts on the opinion prediction tasks of clean and perturbed cases. The Adv-COI shows lower belief truth probability MAE, higher scalability on the large graphs and stronger resistance to strong adversarial attacks, and achieves an improvement of 20~32% prediction accuracy, compared to the best baseline CSL under PGD attack with 0.01 distortion on the opinions. It also outperforms other the state-of-the-art competitive counterparts with high margin on the structure attack experiments.

VIII. Acknowledgement

This work is partially supported by NSF (IIS-1750911, IIS-1815696) and ARL's Competitive Basic Research Program under Computational and Information Sciences Directorate and by the US Army Research Office under grant number W911NF1720129. 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 ARL or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

[1] Reference speed for congestion evaluation. http://www.inrix.com/scorecard/methodology.asp/.
[2] AbouRizk, S. M., Halpin,D. W., and Wilson, J. R. Fitting beta distributions based on sample data. *JCEM*, 120(2):pp.288–305, 1994.
[3] Bach, S.H., Broecheler, M., Huang, B. and Getoor, L. Hinge-loss markov random fields and probabilistic soft logic. *arXiv:1505.04406*, 2015.
[4] Ganchev, K., Gillenwater, J., and Taskar,B. Posterior regularization for structured latent variable models. *JMLR*, Jul,pp.2001-2049, 2010.
[5] Huang, B., Kimmig, A., Getoor L., and Golbeck, J. Probabilistic soft logic for trust analysis in social networks. In *StarAI* , pp. 1-8, 2012.
[6] Jøsang,A. A logic for uncertain probabilities. *J. of Uncertainty, Fuzziness and Knowledge-Based Systems*, 9(03),pp.279-311, 2001.
[7] Jøsang,A. *Subjective Logic: A Formalism for Reasoning Under Uncertainty*. Springer, 2016.
[8] Jøsang,A.,Hayward,R., and Pope, S. Trust network analysis with subjective logic. In *ACSCV*, pp. 85-94, 2006.
[9] Boyd, S., Parikh, N., Chu, E., Peleato, B. and Eckstein, J. Distributed optimization and statistical learning via the alternating direction method of multipliers. *FTML, 3(1)*, pp.1-122, 2011.
[10] Wang, B., Jia, J., Zhang, L. and Gong, N.Z. Structure-based sybil detection in social networks via local rule-based propagation. In *IEEE TNSE*, 2018.
[11] Gong, N.Z., Frank, M. and Mittal, P. Sybilbelief: A semi-supervised learning approach for structure-based sybil detection. *IEEE TIFS*,9(6),pp.976-987, 2014.
[12] Chen, F., Wang, C., and Cho,J.H. Collective subjective logic: Scalable uncertainty-based opinion inference. *In IEEE Big Data* , pp.7-16, 2017.
[13] Huang, B., Kimmig, A., Getoor, L. and Golbeck, J., A flexible framework for probabilistic models of social trust. *In SBP 2013,Springer.*
[14] Chen, P.T., Chen, F. and Qian, Z. Road traffic congestion monitoring in social media with hinge-loss Markov random fields. *IEEE ICDM 2014*.
[15] Zhao, X., Feng C., and Cho J.H. Deep Learning based Scalable Inference of Uncertain Opinions. *IEEE ICDM 2018*.
[16] Liu, Y., Chen X., Liu C., and Song, D.. Delving into transferable adversarial examples and black-box attacks. *In ICLR, 2017*.
[17] Szegedy, C., Zaremba,W. , Sutskever I., Bruna,J., Erhan, D., Goodfellow, I., and Fergus, F. Intriguing properties of neural networks. *In ICLR 2014*.
[18] Kurakin, A., Goodfellow, I., and Bengio, S. Adversarial machine learning at scale. *In ICLR 2017*.
[19] Carlini, N., and Wagner, D. Adversarial examples are not easily detected: Bypassing ten detection methods. *AISec, pp. 3-14. ACM, 2017*.
[20] Goodfellow, I., Shlens,J. , Szegedy, C. Explaining and harnessing adversarial examples. *arXiv:1611.01236 (2016)*.
[21] Huang, S., Papernot, N., Goodfellow, I., Duan, Y. and Abbeel, P. Adversarial attacks on neural network policies. *arXiv:1702.02284 (2017)*.
[22] Dai, H., Li, H., Tian, T., Huang, X., Wang, L., Zhu, J. and Song, L. Adversarial Attack on Graph Structured Data. *PMLR 80,2018*.
[23] Zügner, D., Akbarnejad, A. and Günnemann, S. Adversarial attacks on neural networks for graph data. *SIGKDD, pp. 2847-2856. ACM, 2018.*.
[24] Madry, A., Makelov, A., Schmidt, L., Tsipras, D. and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv:1706.06083 (2017)*.