IDENTIFYING SUSCEPTIBLE AGENTS IN TIME VARYING OPINION DYNAMICS THROUGH COMPRESSIVE MEASUREMENTS

Hoi-To Wai[†]*, Asuman E. Ozdaglar*[‡]*, Anna Scaglione*[†]

[†]School of ECEE, Arizona State University, Tempe, AZ, USA. [‡]LIDS, MIT, Cambridge, MA, USA.

Emails: htwai@asu.edu, asuman@mit.edu, Anna.Scaglione@asu.edu

ABSTRACT

We provide a compressive-measurement based method to detect *susceptible* agents who may receive misinformation through their contact with 'stubborn agents' whose goal is to influence the opinions of agents in the network. We consider a DeGroot-type opinion dynamics model where regular agents revise their opinions by linearly combining their neighbors' opinions, but stubborn agents, while influencing others, do not change their opinions. Our proposed method hinges on estimating the *temporal difference vector* of network-wide opinions, computed at time instances when the stubborn agents interact. We show that this temporal difference vector has approximately the same support as the locations of the susceptible agents. Moreover, both the interaction instances and the temporal difference vector can be estimated from a small number of aggregated opinions. The performance of our method is studied both analytically and empirically. We show that the detection error decreases when the social network is better connected, or when the stubborn agents are 'less talkative'.

***Index Terms*—** opinion dynamics, compressive sensing, stubborn agents, malicious agents, spread of misinformation

1. INTRODUCTION

Online social media platforms such as Twitter, Facebook, are viewed as the next generation source for news in lieu of the traditional TV or newspaper media. In particular, news items are conveyed through the 're-tweeting' or 'sharing' actions and diffused to the other agents on the online social networks (OSNs). Due to its decentralized structure, the OSNs are considered more effective in conveying information than the traditional media. However, it is known that OSNs are plagued by the injection of so-called 'fake news' or 'alternative facts' [1] by a variety of actors, and that false information can spread across the network relatively easily. Needless to say, there is widespread concern on how this can nefariously impact society.

In this paper, we consider a time varying opinion dynamics model to approximate the spread of false information in OSNs, with the goal of identifying agents who may be subject to such manipulations. Our model is closely related to the DeGroot model [2] and randomized gossiping model [3, 4]. Similar to [4], we model the false information sources as *stubborn agents* (a.k.a. *forceful agents*) whose opinions remain constant throughout the consensus process. Under this setting, it is known that the opinions of *all* agents in the social network will be shaped by those of the stubborn agents, resulting in the spread of false information. We combat the spread of false information by identifying the set of *susceptible agents*. The latter are agents who are in *direct* contact with stubborn agents and are the gateway for the stubborn agents to influence the rest of the social network. As observed from Fig. 1, if the links between these susceptible agents and the stubborn agents are severed, we can isolate the stubborn agents and eliminate their influences on the social network. This may be accomplished by incentivizing the susceptible agents to employ fact-checkers (e.g., `http://snopes.com`) before 'sharing' or 'retweeting'.

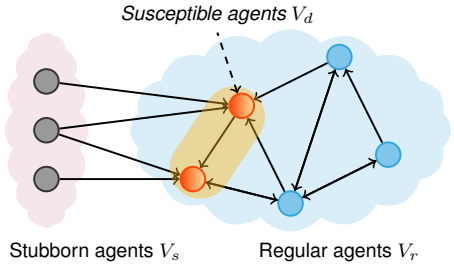

Fig. 1. Illustrating the agents of different roles in the social network.

This paper proposes a compressive measurement-based method for detecting the set of susceptible agents. Our development consists of two parts. First, we demonstrate that the *temporal difference* of the agents' opinions is supported on the locations of the susceptible agents. Second, when the number of susceptible agents is small, we show that the temporal difference vector can be recovered from a small number of measurements through solving a LASSO problem. The set of susceptible agents can then be detected by identifying the support of the recovered vector. We present analytical results with insights on the detection performance. Interestingly, we show that the detection performance improves as the stubborn agents becomes less 'talkative' and the social network between regular agents forms a closer knit, *i.e.,* the regular agents have more links to each other and trusts each other more. Numerical experiments are performed to verify our claims. Our results shed light into how to *engineer* a social network that is resilient to false information campaigns.

Related Work. Our work is related to a number of recent work on data falsification detection in consensus networks [5–8]. Under a wireless sensor network scenario, these work focus on designing consensus protocols that are resilient to the data falsification attacks launched by stubborn agents except for our previous work in [6] which proposed a detection method for identifying stubborn agents. We focus on the OSN scenario where the consensus protocol is viewed as a fixed dynamics rule in the social network. Moreover, this paper is related to the sparsity-based network anomaly detection method in [9, 10]. In comparison, we provide analysis that highlights on the properties of a *robust* social network.

Last, it is worthwhile to mention that alternative models of attacks on social networks opinion dynamics and of defense mechanisms against such attacks can be found in [11–13].

This work is supported by NSF CCF-BSF 1714672.

Notations. For any natural number $n \in \mathbb{N}$, we denote $[n]$ as the set $\{1, 2, ..., n\}$. Vectors (*resp.* matrices) are denoted by boldfaced letters (*resp.* capital letters). We denote x_i or $[\boldsymbol{x}]_i$ as the ith element of the vector $\boldsymbol{x}$. Vector $\mathbf{1}$ is an all-one vector with compatible dimension. The superscript $(\cdot)^\top$ denotes matrix/vector transpose. $\|\cdot\|_2$ is the standard Euclidean norm and $\|\cdot\|_1$ is the ℓ_1-norm.

2. TIME VARYING OPINION DYNAMICS MODEL

Consider a social network described by a time varying, directed graph $G(t) = (V, E(t))$ with $V = [n + S] := \{1, ..., n + S\}$ and $E(t) \subseteq V \times V$ such that $(i, i) \in E(t)$ for all $i \in V$. We use $(i, j) \in E(t)$ to denote an edge *pointing from i to j* in $E(t)$. These agents are divided into two groups — $V_s := [S]$ is the set of stubborn agents and $V_r := V \setminus V_s$ is the set of regular agents. We also define $G := (V, E)$ with $E := \cup_{t=0}^{\infty} E(t)$ to collect all the selected edges. The goal of the stubborn agents is to influence the other agents with their opinions.

Formally, the initial opinion of each regular agent is given by their perceived *state-of-the-world* such that $x_i(0) = \theta_i$. While each stubborn agent holds an initial opinion $x_i(0) = \alpha_i$ which does not change over time. Let $\boldsymbol{x}(t) \in \mathbb{R}^{n+S}$ be the vector of opinions of agents at time $t \geq 0$. Similar to the DeGroot model [2], the opinions evolve as:

$$\boldsymbol{x}(t+1) = \boldsymbol{W}(t)\boldsymbol{x}(t) \text{ where } \boldsymbol{W}(t) = \begin{pmatrix} \boldsymbol{I} & \mathbf{0} \\ \boldsymbol{B}(t) & \boldsymbol{D}(t) \end{pmatrix} , \quad (1)$$

where the sub-matrices $\boldsymbol{B}(t) \in \mathbb{R}_+^{n \times S}$, $\boldsymbol{D}(t) \in \mathbb{R}_+^{n \times n}$ correspond to the sub-graphs between stubborn and regular agents, regular and regular agents, respectively. Note that $\boldsymbol{W}(t)$ is a weighted adjacency matrix of $G(t)$. If we partition $\boldsymbol{x}(t)$ as $(\boldsymbol{x}_s(t); \boldsymbol{x}_r(t))$ where $\boldsymbol{x}_s(t) \in \mathbb{R}^S$ and $\boldsymbol{x}_r(t) \in \mathbb{R}^n$ denote the opinions of the stubborn and regular agents, respectively, then the above dynamics enforces $\boldsymbol{x}_s(t + 1) = \boldsymbol{x}_s(t)$ for all $t \geq 0$. Our assumptions on $G(t)$, $\boldsymbol{W}(t)$ are:

H1. *For all $t \geq 0$, if $(i, j) \in E(t)$, then $W_{ji}(t) \geq \eta > 0$.*

H2. *For all $t \geq 0$, it holds that $\boldsymbol{W}(t) \geq \mathbf{0}$ and $\boldsymbol{W}(t)\mathbf{1} = \mathbf{1}$.*

H3. *There exists an integer C_1 such that for all $t \geq 0$, the sub-graph $(V_r, E(V_r; t) \cup \ldots \cup E(V_r; t + C_1 - 1))$ is strongly connected. Note that $E(V_r; t)$ refers to the edge set $E(t)$ restricted to V_r.*

H4. *There exists an integer C_2 such that for all $t \geq 0$, there exists a pair $i \in V_r$, $j \in V_s$ with $(j, i) \in E(t) \cup \ldots \cup E(t + C_2 - 1)$.*

These assumptions are rather standard. H3 says that the regular network is strongly connected infinitely often and H4 says that the stubborn agents influence at least a regular agent infinitely often. Next, we define

$$\mathcal{T}_{\mathsf{stub}} := \{t \geq 0 : [\boldsymbol{B}(t)]_{ij} \neq 0, \text{ for some } i, j\} \quad (2)$$

as the time instances when a stubborn agent is active. Moreover,

H 5. *For $t \notin \mathcal{T}_{\mathsf{stub}}$, the matrix $\boldsymbol{D}(t)$ is doubly stochastic with $\boldsymbol{D}(t)\mathbf{1} = \boldsymbol{D}(t)^\top \mathbf{1} = \mathbf{1}$.*

H5 states that when the stubborn agents are inactive, the regular agents always attempt to compute the average $\bar{\theta} = (1/n)\sum_{i=1}^n \theta_i$. We remark that $G(t), \boldsymbol{W}(t)$ satisfying H1 to H5 are closely related to the time varying graphs taken from a randomized gossiping model [3]. Moreover, the opinion dynamics is parameterized by:

$$L := \max_{\ell \in \mathbb{Z}} \ \ell \ \text{ s.t. } \ \ell \leq |t_i - t_j|, \ \forall\, t_i, t_j \in \mathcal{T}_{\mathsf{stub}}, \ t_i \neq t_j \ , \quad (3)$$

where $1 \leq L \leq C_2$ captures the 'talkativeness' of the stubborn agents, *i.e.,* if L decreases, the stubborn agents influence others more frequently and therefore 'more talkative'.

Finally, let us comment on the asymptotic opinions resulting from the opinion dynamics (1). There are two cases — when the stubborn agents are coordinated, *i.e.,* $\boldsymbol{x}_s(0) = \alpha\mathbf{1}$, then it can be shown that $\lim_{t\to\infty} \boldsymbol{x}(t) = \alpha\mathbf{1}$; when the stubborn agents are not coordinated, *i.e.,* $\boldsymbol{x}_s(0) \notin \text{span}\{\mathbf{1}\}$, then the sequence $\{\boldsymbol{x}(t)\}_{t=1}^{\infty}$ may not converge and the opinions may fluctuate. These observations are similar to those proven in [14] for a randomized gossiping model.

Our goal is to prevent the stubborn agents from spreading false information in the social network. To do so, we aim to identify the set of *susceptible agents*, who are the regular agents in direct contact with at least one stubborn agent [cf. Fig. 1], *i.e.,* the set

$$V_d := \{i \in V_r : (j, i) \in E, \text{ for some } j \in V_s\} \ . \quad (4)$$

Note we can equivalently write as $V_d = \{i + S : [\sum_{t=0}^{\infty} \boldsymbol{B}(t)\mathbf{1}]_i > 0\}$, *i.e.,* the support of the non-negative vector $\sum_{t=0}^{\infty} \boldsymbol{B}(t)\mathbf{1}$. Hence detecting the susceptible agents can be cast as one of identifying the support of the sum vector $\sum_{t=0}^{\infty} \boldsymbol{B}(t)\mathbf{1}$.

3. DETECTING THE SUSCEPTIBLE AGENTS

We propose to detect the susceptible agents through observing the transient states of the opinion dynamics. In this section, we first analyze a temporal difference vector that reveals the locations of V_d, then we present an estimation method which identifies V_d from compressed measurements of the transient opinions.

3.1. Temporal Difference Vector

As the stubborn agents' opinions must propagate through the susceptible agents before reaching the rest of the social network, it is anticipiated that the *temporal difference vector*, defined as

$$\Delta\boldsymbol{x}(t) := \boldsymbol{x}(t + 1) - \boldsymbol{x}(t) \ , \quad (5)$$

contain large values (or 'spikes') at the indices V_d when $t \in \mathcal{T}_{\mathsf{stub}}$ since stubborn agents' opinions are typically different from the rest. The intuition stems from the fact that $\Delta\boldsymbol{x}(t)$ records the immediate *impact* experienced by the susceptible agents due to the active stubborn agents. To observe this, we rewrite the regular agents' opinions as:

$$\boldsymbol{x}_r(t + 1) = \boldsymbol{D}(t, 0)\boldsymbol{x}_r(0) + \tilde{\boldsymbol{x}}_r(t) \ , \quad (6)$$

where

$$\tilde{\boldsymbol{x}}_r(t) := \boldsymbol{B}(t)\boldsymbol{x}_s(0) + \textstyle\sum_{s=0}^{t-1} \boldsymbol{D}(t, s + 1)\boldsymbol{B}(s)\boldsymbol{x}_s(0) \ , \quad (7)$$

and

$$\boldsymbol{D}(t, s) := \begin{cases} \boldsymbol{D}(t)\boldsymbol{D}(t - 1)\cdots\boldsymbol{D}(s), & \text{if } t \geq s \ , \\ \boldsymbol{I}, & \text{if } s > t \ . \end{cases} \quad (8)$$

Now, if we partition $\Delta\boldsymbol{x}(t) = (\Delta\boldsymbol{x}_s(t); \Delta\boldsymbol{x}_r(t))$ in a similar fashion as $\boldsymbol{x}(t)$, it can be shown that:

$$\begin{pmatrix} \Delta\boldsymbol{x}_s(t) \\ \Delta\boldsymbol{x}_r(t) \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \boldsymbol{B}(t)\boldsymbol{x}_s(0) + (\boldsymbol{D}(t) - \boldsymbol{I})\tilde{\boldsymbol{w}}(t) \end{pmatrix} , \quad (9)$$

where $\tilde{\boldsymbol{w}}(t) = \tilde{\boldsymbol{x}}_r(t - 1) + \boldsymbol{D}(t - 1, 0)\boldsymbol{x}_r(0)$ is a residual term related to the opinion mixing before the impact from stubborn agents

at $t \in \mathcal{T}_{\mathsf{stub}}$. Note that $\Delta \boldsymbol{x}_s(t)$ is always zero as the opinions of the stubborn agents are constant.

When $t \in \mathcal{T}_{\mathsf{stub}}$, the first term $\boldsymbol{B}(t)\boldsymbol{x}_s(0)$ in (9) is supported on the set of active susceptible agents as desired. Moreover, the residual's magnitude is controlled by the parameters C_1 and L:

Proposition 1. *If the opinion dynamics is initialized with $\boldsymbol{x}_s(0) \geq \mathbf{0}$, then for all $t \in \mathcal{T}_{\mathsf{stub}}$, we have*

$$\begin{aligned}\Delta \boldsymbol{x}_r(t) = \ & (1/n)\|\alpha \mathbf{1} - \tilde{\boldsymbol{x}}_r(t-L-1)\|_1 \Big(\boldsymbol{B}(t)\mathbf{1} + \boldsymbol{w}(t)\Big) \\ & - \boldsymbol{B}(t)\boldsymbol{x}_s^0 + (\boldsymbol{D}(t) - \boldsymbol{I})\boldsymbol{D}(t-1,0)\boldsymbol{x}_r(0) \ ,\end{aligned} \tag{10}$$

where $\alpha := \|\boldsymbol{x}_s(0)\|_\infty$, $\boldsymbol{x}_s^0 := \alpha \mathbf{1} - \boldsymbol{x}_s(0) \geq \mathbf{0}$ and

$$\|\boldsymbol{w}(t)\|_1 \leq n^2 \cdot M \cdot \lambda^L \tag{11}$$

such that

$$M := 2 \cdot \frac{1 + \eta^{-(n-1)C_1}}{1 - \eta^{(n-1)C_1}}, \ \lambda := (1 - \eta^{(n-1)C_1})^{\frac{1}{(n-1)C_1}} . \tag{12}$$

The proof involves a careful analysis of (9) with the given assumptions and bounding the mixture of doubly stochastic matrices as in [15], the details can be found in an online appendix.[1] Note that the condition $\boldsymbol{x}_s(0) \geq \mathbf{0}$ can be satisfied without loss of generality if we consider an equivalent opinion dynamics with the initializations $\boldsymbol{x}'_s(0) = \boldsymbol{x}_s(0) + \beta \mathbf{1}, \boldsymbol{x}'_r(0) = \boldsymbol{x}_r(0) + \beta \mathbf{1}$ for some large $\beta > 0$.

Now, suppose that $\boldsymbol{x}_s^0, \boldsymbol{x}_r(0) \approx \mathbf{0}$, from the proposition the magnitude of $\Delta \boldsymbol{x}_r(t)$ is controlled by the envelope $\frac{1}{n}\|\alpha \mathbf{1} - \tilde{\boldsymbol{x}}_r(t - L - 1)\|_1$. In addition, $\Delta \boldsymbol{x}_r(t)$ has a signal component $\boldsymbol{B}(t)\mathbf{1}$ and a residual component $\boldsymbol{w}(t)$. The residual $\boldsymbol{w}(t)$ can be reduced when:

- the regular-regular network is well connected, *i.e.,* the joint connectivity constant C_1 is small;
- the stubborn agents are not 'talkative', *i.e.,* the talkativeness parameter L is large [cf. (3)].

Intuitively, if the above conditions hold and we fix $t \in \mathcal{T}_{\mathsf{stub}}$, then the set of social network's opinions is in a neighborhood of the steady state at $t-1$, therefore any action of the stubborn agents applies a 'strong reset' to the social network, causing a large spike on the temporal difference vector. Crucially, these conditions give insight into how a 'robust' social network should behave, where the spread of false information can be easily detected from $\Delta \boldsymbol{x}(t)$. Lastly, we remark that our bound is based on a worst case analysis. It is loose in general as it requires $L \gg (n-1)C_1$ to guarantee a low noise power in $\boldsymbol{w}(t)$, yet the numerical experiments in Section 4 shows far better performance than the bound.

3.2. Identifying V_d with Compressive Measurements

Proposition 1 suggests that when $t \in \mathcal{T}_{\mathsf{stub}}$ and $\|\boldsymbol{x}_s(0)\|_\infty - \boldsymbol{x}_s(0), \boldsymbol{x}_r(0)$ are small, the active susceptible agents can be revealed through detecting the locations of spikes in the temporal difference vector $\Delta \boldsymbol{x}(t)$. Similarly, the set V_d can be revealed through detecting the locations of spikes in the sum vector $\sum_{t \in \mathcal{T}_{\mathsf{stub}}} \Delta \boldsymbol{x}(t)$. However, computing $\Delta \boldsymbol{x}(t)$ requires accruing the opinions of *all* agents, which may not be practical or possible; moreover, the set of time instances when the stubborn agents are active is unknown.

To overcome the hurdles above, we assume that the number of susceptible agents is small with $|V_d| \ll n$. This assumption can be justified since the false information sources (stubborn agents) are not mainstream sources and hence are typically in direct contact with a few agents. Consequently, the temporal difference vector $\Delta \boldsymbol{x}(t)$ is the combination of a *sparse vector* supported on the active susceptible agents and a noise vector. This inspires us to apply a compressive sensing approach as follows.

Let us begin by formally describing our observation model. Over an observation period from time T_0 to $T_{\mathsf{max}} + 1$, we observe m linear measurements of the opinions with $m \ll n + S$:

$$\boldsymbol{y}(t) = \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{z}(t), \ t \in \{T_0, ..., T_{\mathsf{max}} + 1\} \ , \tag{13}$$

where $\boldsymbol{A} \in \mathbb{R}^{m \times (n+S)}$ is a known measurement matrix and $\boldsymbol{z}(t)$ is a zero-mean, additive noise. In practice, the observations $\boldsymbol{y}(t)$ can be obtained by surveying m different groups of agents on the OSN and the noise $\boldsymbol{z}(t)$ models the error of estimating the aggregated opinions of each group. As $m \ll n+S$, we do not need to exhaustively survey the opinions of the agents one by one. Furthermore, we assume the following on $\boldsymbol{A}$:

H 6. *The all-one vector is in the row span of $\boldsymbol{A}$. In other words, there exists $\boldsymbol{c} \in \mathbb{R}^m$ such that $\mathbf{1}^\top = \boldsymbol{c}^\top \boldsymbol{A}$.*

H6 implies that every agent's opinions will be represented in the aggregated group opinions given by $\boldsymbol{y}(t)$.

Detecting $\mathcal{T}_{\mathsf{stub}} \cap [T_0, T_{\mathsf{max}}]$ from the observations. Under H6, it is possible to detect the instances within the observation period in which the stubborn agents are active, *i.e.,* the set $\mathcal{T}_{\mathsf{stub}} \cap [T_0, T_{\mathsf{max}}]$. To see this, define $\Delta \boldsymbol{y}(t) := \boldsymbol{y}(t+1) - \boldsymbol{y}(t)$ and consider:

$$\begin{aligned}\boldsymbol{c}^\top \Delta \boldsymbol{y}(t) &= \boldsymbol{c}^\top \boldsymbol{A} \Delta \boldsymbol{x}(t) + \boldsymbol{c}^\top (\boldsymbol{z}(t+1) - \boldsymbol{z}(t)) \\ &= \mathbf{1}^\top \Delta \boldsymbol{x}(t) + \boldsymbol{c}^\top (\boldsymbol{z}(t+1) - \boldsymbol{z}(t)) \ .\end{aligned} \tag{14}$$

When $t \notin \mathcal{T}_{\mathsf{stub}}$, we have:

$$\mathbf{1}^\top \Delta \boldsymbol{x}(t) = \mathbf{1}^\top (\boldsymbol{W}(t) - \boldsymbol{I})\boldsymbol{x}(t) = 0 \ , \tag{15}$$

since the $\boldsymbol{W}(t)$ in the above is a block diagonal matrix composed of $\boldsymbol{I}$ and $\boldsymbol{D}(t)$, both matrices are doubly stochastic [cf. H5]. When $t \in \mathcal{T}_{\mathsf{stub}}$, we have $\mathbf{1}^\top \Delta \boldsymbol{x}(t) \neq 0$ in general. We can show:

Proposition 2. *If the noise $\boldsymbol{z}(t)$ is i.i.d. and sub-Gaussian such that $\mathbb{E}[\exp(s \cdot \boldsymbol{c}^\top \boldsymbol{z}(t)] \leq \exp(\sigma_z^2 s^2/2)$ for all $s \in \mathbb{R}$, then the false-alarm and missed detection rates are given as:*

$$P\big(|\boldsymbol{c}^\top \Delta \boldsymbol{y}(t)| > \delta | t \notin \mathcal{T}_{\mathsf{stub}}\big) \leq 2 \cdot \exp(-\delta^2/(4\sigma_z^2)) \ , \tag{16}$$

$$P\big(|\boldsymbol{c}^\top \Delta \boldsymbol{y}(t)| \leq \delta | t \in \mathcal{T}_{\mathsf{stub}}\big) \leq \exp(-(\delta - |m_t|)^2/(4\sigma_z^2)) \ , \tag{17}$$

where $m_t := \mathbf{1}^\top \Delta \boldsymbol{x}(t)$ is the sum of the temporal difference.

The proof, which is based on the standard Chernoff bound of sub-Gaussian random variables, can be found in the online appendix.[1]

The above proposition suggests that $\mathcal{T}_{\mathsf{stub}} \cap [T_0, T_{\mathsf{max}}]$ can be estimated by thresholding on $|\boldsymbol{c}^\top \Delta \boldsymbol{y}(t)|$. Let $\delta > 0$, we set

$$\hat{\mathcal{T}}_{\mathsf{stub}} = \{t \in [T_0, T_{\mathsf{max}}] : |\boldsymbol{c}^\top \Delta \boldsymbol{y}(t)| > \delta\} \ . \tag{18}$$

We observe that the detection performance depends on the noise variance σ_z^2 and the magnitude of m_t. Let us comment on $|m_t|$ whose analytical expression can be found in the online appendix.[1] We consider the special case when $\boldsymbol{x}_s(0) = \alpha \mathbf{1}$. As $\lim_{t \to \infty} \boldsymbol{x}(t) = \alpha \mathbf{1}$, we have $m_t \to 0$. It implies that while a large T_{max} allows for including more samples to detect V_d, the detection performance for $\mathcal{T}_{\mathsf{stub}}$ may be degraded as the missed detection rate in (17) increases with t. This suggests a tradeoff in designing the observation period.

[1] `http://www.public.asu.edu/~hwai2/pdf/sus_app.pdf`.

Identifying the set of susceptible agents. Now suppose that $\hat{\mathcal{T}}_{\mathsf{stub}} \approx \mathcal{T}_{\mathsf{stub}} \cap [T_0, T_{\mathsf{max}}]$, we apply Proposition 1 and observe that:

$$\sum_{t\in\hat{\mathcal{T}}_{\mathsf{stub}}} \Delta \boldsymbol{y}(t) \approx \boldsymbol{A} \sum_{t\in\hat{\mathcal{T}}_{\mathsf{stub}}} \Big(\boldsymbol{B}(t)(\mathbf{1}+\tilde{\boldsymbol{x}}_s^0) + \tilde{\boldsymbol{w}}(t)\Big) + \tilde{\boldsymbol{z}}(t)\ , \quad (19)$$

where $\tilde{\boldsymbol{w}}(t), \tilde{\boldsymbol{z}}(t)$ are additive noise depending on $\boldsymbol{x}_r(0), \boldsymbol{z}(t)$ and $\tilde{\boldsymbol{x}}_s^0$ is some vector that depends on $\boldsymbol{x}_s^0$. In the above, $\boldsymbol{b} = \sum_{t\in\hat{\mathcal{T}}_{\mathsf{stub}}} \boldsymbol{B}(t)\mathbf{1}$ is the desired vector as it is supported on the locations of the susceptible agents. Since $|V_d| \ll n$, the vector $\boldsymbol{b}$ is sparse. Naturally, we can recover $\boldsymbol{b}$ with the LASSO problem:

$$\boldsymbol{b}^\star = \arg\min_{\tilde{\boldsymbol{b}}\in\mathbb{R}^{n+S}} \Big\| \frac{1}{|\hat{\mathcal{T}}_{\mathsf{stub}}|} \sum_{t\in\hat{\mathcal{T}}_{\mathsf{stub}}} \Delta\boldsymbol{y}(t) - \boldsymbol{A}\tilde{\boldsymbol{b}} \Big\|_2^2 + \rho\|\tilde{\boldsymbol{b}}\|_1\ , \quad (20)$$

where $\rho > 0$ is a regularization parameter. Finally, we detect the support of $\boldsymbol{b}^\star$, *i.e.,* let $\epsilon > 0$ be a predefined threshold, by

$$\hat{V}_d = \{i \in [n+S] : |[\boldsymbol{b}^\star]_i| > \epsilon\}\ . \quad (21)$$

The performance analysis of the LASSO detector is left for future work. We remark that its performance depends on the design of $\boldsymbol{A}$ and can be analyzed under the framework of [16]. In particular, we expect a reasonable support recovery performance if $m = \Omega(|V_d|)$, followed from standard compressive sensing theory.

4. NUMERICAL EXPERIMENTS

We present numerical results on the detection performance for susceptible agents using synthetic data. We consider a network with $n = 200$ regular agents and $S = 20$ stubborn agents and evaluate the performance running 1000 Monte-Carlo trials. The subgraph of regular agents, $G[V_r]$, is a random Erdos-Renyi (ER) graph with connectivity of $2\log n/n$, while the graph that connects the stubborn-regular agents is generated by randomly choosing $|V_d| = 10$ regular agents in V_r and connecting them to 3 random stubborn agents. We initialize the opinions as $\boldsymbol{x}_s(0) = 10\cdot\mathbf{1}$ and $\boldsymbol{x}_r(0) \sim \mathcal{N}(\mathbf{0}, 0.5\mathbf{I})$. Emulating the design proposed in [17] which generates an expander bipartite graph, the sensing matrix $\boldsymbol{A}$ has $d = 5$ random entries of each column equal to 1, and the rest are 0. The measurement noise is $\boldsymbol{z}(t) \sim \mathcal{N}(\mathbf{0}, 0.1\boldsymbol{I})$. The observations $\boldsymbol{y}(t)$ emulate a set of m surveys taking the average opinions of agents. We set $\delta = 1$ for the stubborn agents' activity detection in (18) and $\rho = 0.5$ for (20). The performance is measured by the area under the ROC curve (AUROC) and the area under precision-recall curve (AUPR), computed by sweeping a range of values for ϵ in (21). A perfect detection is attained when AUROC = AUPR = 1.

We consider a randomized gossip protocol [3] with the following modifications. At each time t, a regular agent is selected with probability $(1-\gamma)/|V_r|$ while a stubborn agent is selected with probability $\gamma/|V_s|$. Intuitively, reducing γ increases the constant L in (3) as the stubborn agents become less 'talkative'. When a regular agent is selected, he/she exchanges opinion with R randomly chosen neighboring regular agents with an influence weight of $1/(R+1)$. Increasing R effectively reduces the constant C_1 in H3 as the resulting (union of) time varying graph(s) contains more edges, *i.e.,* the regular agents are better connected with a large R. Lastly, we choose $T_0 = 5\times 10^3/R$ and $T_{\mathsf{max}} = 5\times 10^4/R$.

In the first example in Fig. 2 we study the detection performance of susceptible agents with different degrees of connectivity for the regular-regular social network, by varying R. We compare (1-AUROC) and (1-AUPR) against the number of measurements, m, made per snapshot of the opinions. First, we observe from Fig. 2

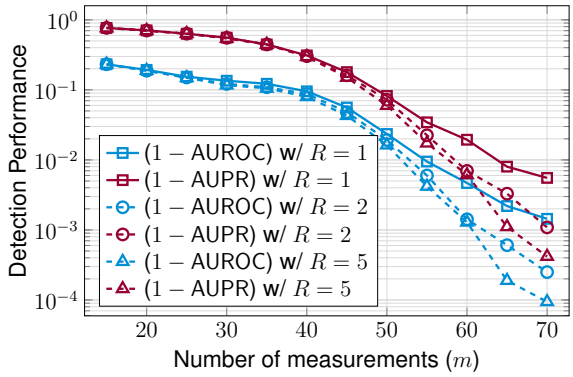

Fig. 2. **(Detection performance with varying connectivity of regular-regular social network R)**. Detection performance against the number of measurements m made on each network snapshot. Larger R implies better connectivity. The gossiping's parameter α is fixed at $|V_s|/|V_r|$, *i.e.,* all agents are equally talkative.

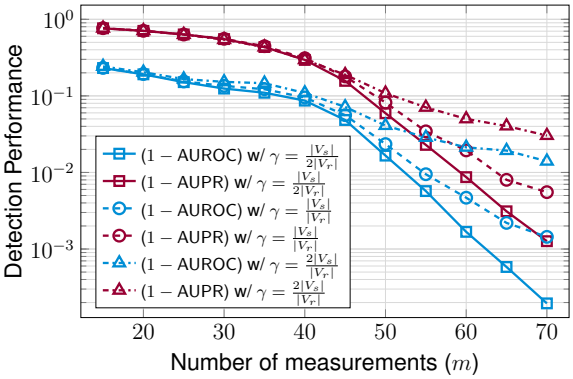

Fig. 3. **(Detection performance with varying stubborn agents' talkativeness level α)**. Detection performance against the number of measurements m made on each network snapshot. The gossiping's parameter is set as $R = 1$, *i.e.,* the pairwise gossip exchange.

that as the number of measurement increases, both AUROC and AUPR approach 1, indicating that the detection performance improves with larger m, and it is reasonably good when $m \approx 50$. Second, the detection performance improves when R increases, corroborating our theoretical claim in Proposition 1. This shows that a well connected social network can be better defended.

The second numerical example examines the effects of talkativeness of the stubborn agents [cf. (3)] by varying γ. Fig. 3 shows the comparison of detection performance. As observed from the figure, the detection performance worsens when γ increases, *i.e.,* the stubborn agents becomes more 'talkative'. This corroborates with our analysis in Proposition 1, which predicts that the detection performance degrades as L decreases. Nevertheless, one can mitigate this loss with a sufficient number of measurements.

Conclusions. In this paper, we proposed a detection method for agents that are exposed directly to stubborn influencers. Isolating these susceptible agents is essential to defend a social network against the influences of false information. Under the assumption that the group is small in size, we show that the number of measurements to monitor can be far less than the number of agents in the network. Analytical results give insights on the detection performance, which is shown to depend on the connectivity of the network and the talkativeness of stubborn agents. Our results are verified through numerical experiments.

5. REFERENCES

[1] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," National Bureau of Economic Research, Tech. Rep., 2017.

[2] M. DeGroot, "Reaching a consensus," in *Journal of American Statistcal Association*, vol. 69, 1974, pp. 118–121.

[3] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Randomized gossip algorithms," *IEEE/ACM Transactions on Networking (TON)*, vol. 14, no. SI, pp. 2508–2530, 2006.

[4] D. Acemoglu, A. Ozdaglar, and A. ParandehGheibi, "Spread of (mis) information in social networks," *Games and Economic Behavior*, vol. 70, no. 2, pp. 194–227, 2010.

[5] L. Su and N. Vaidya, "Byzantine multi-agent optimization: Part i," *arXiv preprint arXiv:1506.04681*, 2015.

[6] R. Gentz, S. X. Wu, H.-T. Wai, A. Scaglione, and A. Leshem, "Data injection attacks in randomized gossiping," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 2, no. 4, pp. 523–538, 2016.

[7] S. Sundaram and B. Gharesifard, "Consensus-based distributed optimization with malicious nodes," in *Communication, Control, and Computing (Allerton), 2015 53rd Annual Allerton Conference on*. IEEE, 2015, pp. 244–249.

[8] B. Kailkhura, S. Brahma, and P. K. Varshney, "Data falsification attacks on consensus-based detection systems," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 3, no. 1, pp. 145–158, 2017.

[9] G. Mateos and K. Rajawat, "Dynamic network cartography: Advances in network health monitoring," *IEEE Signal Processing Magazine*, vol. 30, no. 3, pp. 129–143, 2013.

[10] M. Mardani, G. Mateos, and G. B. Giannakis, "Dynamic anomalography: Tracking network anomalies via sparsity and low rank," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 1, pp. 50–66, 2013.

[11] A. Beutel, W. Xu, V. Guruswami, C. Palow, and C. Faloutsos, "Copycatch: stopping group attacks by spotting lockstep behavior in social networks," in *Proceedings of the 22nd international conference on World Wide Web*. ACM, 2013, pp. 119–130.

[12] Q. Cao, X. Yang, J. Yu, and C. Palow, "Uncovering large groups of active malicious accounts in online social networks," in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2014, pp. 477–488.

[13] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, "The rise of social bots," *Communications of the ACM*, vol. 59, no. 7, pp. 96–104, 2016.

[14] D. Acemoğlu, G. Como, F. Fagnani, and A. Ozdaglar, "Opinion fluctuations and disagreement in social networks," *Mathematics of Operations Research*, vol. 38, no. 1, pp. 1–27, 2013.

[15] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Transactions on Automatic Control*, vol. 54, no. 1, pp. 48–61, 2009.

[16] S. Foucart and H. Rauhut, *A mathematical introduction to compressive sensing*. Birkhäuser Basel, 2013, vol. 1, no. 3.

[17] S. Jafarpour, W. Xu, B. Hassibi, and R. Calderbank, "Efficient and robust compressed sensing using optimized expander graphs," *IEEE Transactions on Information Theory*, vol. 55, no. 9, pp. 4299–4308, 2009.