Trust in Multi-Agent Networks: From Self-Centered to Team-Oriented

Tina Setter Andrea Gasparri Magnus Egerstedt

Abstract— **In this work, we incorporate trust into the interaction dynamics for multi-agent systems in order to analyze the effects of trust on human-robot teams. We extend previous work by moving away from a "self-centered" trust model and introduce a so-called "team-oriented" trust model in which each agent increases its trust for its neighbors only if they are collectively contributing to the team's overall goal. The coupling of the trust dynamics and the agents' state dynamics is shown to give rise to an intricate relationship that has the potential to make the team's performance deteriorate under certain circumstances. We derive conditions under which the multi-agent system is guaranteed to achieve a collective objective and provide simulations to corroborate the theoretical findings.**

I. Introduction

The area of multi-agent robotics has grown in popularity in the last decade, as the idea of deploying many simple robots as opposed to one more complex robot has far-reaching benefits and applications. One main benefit of multi-agent systems is that they are robust to failures; that is, if one robot fails, there is an abundance of robots remaining to complete the task. Additionally, these robots can often be made smaller, less complex, and thus ultimately less expensive at an individual module level than one large, complex robot. Many applications for multi-agent systems have been alluded to in the literature, including space exploration [1], military missions [2], and search and rescue [3], to name a few.

Robots are becoming more increasingly embedded into our everyday lives, and therefore humans and robots working together is becoming a reality, as can already be seen in industrial applications such as manufacturing [4] and in self-driving cars, for instance. Unsurprisingly, the area of human-robot interaction (HRI) has gained significant attention since the development of these new technologies. Research in this area spans from the analysis of HRI through user-studies [5] to the design of robots that allow humans to be more comfortable around them [6] to the development of metrics for analyzing human-robot systems [7].

In human-robot interactions, it has been shown that human trust in its robotic partner plays an important role because a lack of trust may make people less willing to accept information provided by the robot and thus will not benefit from the advantages that are typically present in a robotic system [8]. Successful HRI relies on creating appropriate levels of trust, which has been shown to be challenging to ensure [9] and much attention has been made to determining the factors that affect trust in HRI [10].Although the research in this area has mostly focused on one human-one robot interaction scenarios, one can imagine that humans will soon need to also interact with swarms of robots and thus be embedded into these aforementioned multi-agent systems.

Human-swarm interaction (HSI) is becoming increasingly important and is accordingly receiving greater attention in the past decade (see [11] for a survey of the literature), and while studies have shown experimentally what these interactions might look like for specific applications or using specific interaction modalities, e.g. [12], [13], we are far from an all-encompassing theory of HSI. And, in order to make progress towards this theory, work must be done on modeling factors that affect the interactions that occur when humans are injected into multi-robot teams.

As trust has been shown to play a large role in human-robot interactions, it is clear that this importance extends to human-swarm interactions [14], [15]. To this point, however, the work done on analyzing trust in human-swarm interactions has been limited and in the work that does exist, the human is often viewed as an operator. For example, in [16], trust is used to schedule the attention of a human operator between multiple robots and in [17], trust is used to blend commands from a human operator with those from an autonomous controller when teleoperating mobile robots. Instead of viewing the human as an operator, we are interested in human-robot teaming where the human is considered an agent, just like the robots in the team, and where there may in fact be multiple humans.

Trust within groups of people has been widely studied in the psychology and sociology fields. It has been argued that trust is important for the performance of teams within organizations [18], it helps save on transaction costs, and has economic implications [19]. In [20], trust is described as "the degree to which an individual believes that a relationship partner will assist in attaining a specific interdependent goal." In this regard, the purpose of our work is to develop a model for the evolution of trust in human-swarm interaction scenarios and incorporate these trust metrics into the standard multi-agent control algorithms to analyze the effect that trust may have on the performance of human-robot teams.

In [21], we made a first attempt to derive a mathematical model of this trust notion by letting an agent's trust evolve according to how much it benefits from the direct

Tina Setter and Magnus Egerstedt are with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA, Email: {tsetter3,magnus}@gatech.edu.

Andrea Gasparri is with the Department of Engineering, University "Roma Tre", 00146 Rome, Italy. Email: {gasparri}@dia.uniroma3.it

The work by the first and third authors was sponsored by Grant No. N00014-15-1-2115 from the U.S. Office of Naval Research.

interactions with its neighbors. However, as will be shown through an illustrative example, the lack of a direct benefit to an agent does not necessarily imply that its neighbors are not contributing to the overall team objective. Hence, this (partial) knowledge may not always allow an agent to make a correct assessment of the trustworthiness of its neighbors and we therefore refer to this trust model as "self-centered".

In this work, we aim to remedy this by proposing a revised trust model that allows human agents to collect further information about each of its neighbors, i.e., their interactions with their respective neighbors, and we call this "team-oriented" trust. In particular, we show that this information is sufficient in allowing an agent to properly evaluate whether its neighbors are contributing to the overall team goal and thus modify its trust accordingly.

The outline of the paper is as follows. In Section II, we introduce the necessary background literature and preliminaries regarding graph theoretic multi-agent networks and summarize our previous work on trust-based interactions to motivate the new trust modeling. In Section III, the team-oriented trust model is introduced and the corresponding invariance and convergence results are given in Section IV and Section V, respectively. In Section VI, we present simulations that illustrate the findings and concluding remarks and future directions are given in Section VII.

II. Preliminaries

A. Graph Theoretic Multi-Agent Network Modeling

In this paper we focus on multi-agent networks that are defined in a graph theoretic manner, as is often done in the literature, e.g. [22], [23]. We use a static, undirected interaction graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the vertex set $\mathcal{V} = \{v_1, \dots, v_N\}$ contains the nodes or agents in the graph and the static edge set $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is a set of unordered pairs of agents representing links over which agents can exchange information. If $(v_i, v_j) \in \mathcal{E}$, then agents i and j can share information with each other. Because the network is undirected, $(v_i, v_j) = (v_j, v_i)$.

We denote agent i's neighborhood set as $\mathcal{N}_i$, which is defined to be the set of agents with whom agent i can share information, i.e. $\mathcal{N}_i = \{j | (v_i, v_j) \in \mathcal{E}\}$. In addition, we denote with $|\mathcal{N}_i|$ the cardinality of the neighborhood of agent i, that is the number of neighbors agent i has. A graph $\mathcal{G}$ is said to be connected if, given any two vertices v_i and v_j, there is a path along edges from v_i to v_j. A graph is said to be complete if, given any two vertices v_i and v_j, there exists an edge between them, that is $(v_i, v_j) \in \mathcal{E}$.

B. Self-Centered Trust Modeling: An Illustrative Example

In [21], we presented a trust model where agent i's trust in its neighbors evolves according to

$$\dot{\tau}_i = -\sum_{j \in \mathcal{N}_i} \frac{\partial F_{ij}(\|x_i - x_j\|)}{\partial x_j}^T \dot{x}_j, \tag{1}$$

where $F_{ij}(\|x_i - x_j\|)$ is the inter-agent performance cost that agents i and j collectively aim to minimize. This is a self-centered trust notion in that agent i only cares about whether its neighbors are helping to decrease the portions of the cost that are pertinent to agent i. However, an agent could be instantaneously contributing to decreasing the overall team cost, $F(x)$, but not necessarily contributing to decreasing agent i's portion of the cost at a particular point in time, and this trust model could potentially deem it untrustworthy.

Fig. 1. $N = 3$ line graph with circles representing the initial positions of the agents.

For example, suppose the goal of the agents was to meet, or in other words achieve consensus, by minimizing the performance cost $F(x) = \sum_{(v_i, v_j) \in \mathcal{E}} F_{ij}(\|x_i - x_j\|)$, where

$$F_{ij}(\|x_i - x_j\|) = \frac{1}{2}\|x_i - x_j\|^2 \tag{2}$$

and suppose the initial configuration of the agents is as in Fig. 1, where the lines between agents represent links indicating edges in the graph. Letting x_i be the 1-dimensional position of agent i, the initial states are given by $x_1(0) = 0$, $x_2(0) = 5$, and $x_3(0) = 15$. Assume that all of the trust values initially start positive, at $\tau_i(0) = 1$.

As in [21], the states evolve according to a weighted gradient-descent scheme where the trust values are the weights, which, for the cost in (2), is given by

$$\dot{x}_i = \tau_i \sum_{j \in \mathcal{N}_i} (x_j - x_i)$$

Because agent 2 is further from agent 3 than to agent 1, it will initially move to the right (and away from agent 1). Agent 1 will see this as agent 2 not contributing to minimizing their inter-agent cost F_{12}, and thus agent 1's trust toward agent 2 will decrease, according to the self-centered trust model in (1). However, agent 2 is indeed contributing to minimizing the overall cost, $F(x)$ by moving closer to agent 3. The limitation with the old trust model is that agent 1 does not take into account (or simply it is not aware of) what is going on in agent 2's neighborhood.

Since the goal is, however, to minimize the overall team cost $F(x)$, we need a new trust model that resolves this issue. Instead of using a self-centered trust model, we propose that each agent should look at its neighbors' *overall* contributions to decreases in the cost when updating their trust values, which will give us a "team-oriented" trust model. This requires that agents have 2-hop information, meaning they need information from their neighbors' neighbors to calculate trust. Because we claim that these dynamics represent those of humans, this requirement is reasonable in the sense that humans have less restrictive perception of the environment than robots. We will present this modified trust model in Section III.

III. Team-Oriented Trust Modeling

For the purposes of this model, we assume that all of the agents are "human-like" in that they all are capable of reasoning about trust and thus incorporate a trust metric into their dynamics. Let there be N agents, where $x_i \in \mathbb{R}^d$ is the state of agent i, which could be a position or an opinion for example, and $\tau_i \in \mathbb{R}$ is the trust that agent i has for its neighbors. If we stack the agents' states into a vector, we can represent the collective state as $x = [x_1^T, x_2^T, \ldots, x_N^T]^T \in \mathbb{R}^{Nd}$ and the collective trust as $\tau = [\tau_1, \ldots, \tau_N]^T \in \mathbb{R}^N$.

In this work, we consider scenarios in which the desired network performance is specified by the performance cost function $F : \mathbb{R}^{Nd} \to \mathbb{R}$, defined by

$$F(x) = \sum_{(v_i,v_j)\in\mathcal{E}} F_{ij}(\|x_i - x_j\|) \tag{3}$$

where $F_{ij}(\|x_i - x_j\|)$ is the symmetric, pairwise performance cost, as is often done in the multi-agent network literature, for example, for formation control [24]. The goal of the team of agents is to achieve a configuration that minimizes $F(x)$. As is the case for many multi-robot applications, $F(x)$ is not necessarily convex and in general, we can only expect to find local minima.

As motivated by the DeGroot model [25] and done in our previous work in [21], we let the state dynamics of agent i evolve by weighting the standard gradient-descent approach by the trust that agent i has for its neighbors, that is

$$\dot{x}_i = -\tau_i \sum_{j\in\mathcal{N}_i} \frac{\partial F_{ij}(\|x_i - x_j\|)}{\partial x_i}. \tag{4}$$

Note that we do not claim that human behavior is exactly modeled in this way, but that in a human-swarm system, the human would be willing to follow these dynamics. However, because of human cognition, the human may lose (or gain) faith in the system, causing the trust to decrease (or increase) and thus the human behavior to change accordingly.

However, the difference between the model in [21] and the one we introduce here is reflected in the trust dynamics, as discussed in Section II. In this work we focus on "neighborhood" trust, meaning that agent i trusts all of its neighbors the same amount. Because we are concerned with how much each neighbor contributes to decreasing the overall cost, we note that the change in cost is given by

$$\frac{dF(x)}{dt} = \sum_{i=1}^{N} \frac{\partial F(x)}{\partial x_i}^T \dot{x}_i$$

and therefore the direct contribution of agent i to the *decrease* in cost, due to the movement of agent i, is $-\frac{\partial F(x)}{\partial x_i}^T \dot{x}_i$. Thus, the higher this term is, the more that agent i is contributing to the team goal, and thus the more that its neighbors should trust it.

To reflect this, we let the evolution of trust for agent i be the sum of its neighbors' contributions to the decrease in cost,

$$\dot{\tau}_i = -\sum_{j\in\mathcal{N}_i} \frac{1}{|\mathcal{N}_j|} \frac{\partial F(x)}{\partial x_j}^T \dot{x}_j \tag{5}$$

or alternatively,

$$\dot{\tau}_i = -\sum_{j\in\mathcal{N}_i} \left[\frac{1}{|\mathcal{N}_j|} \left(\sum_{k\in\mathcal{N}_j} \frac{\partial F_{kj}(\|x_k - x_j\|)}{\partial x_j} \right)^T \dot{x}_j \right] \tag{6}$$

where we added a scale factor of $1/|\mathcal{N}_j|$ to the contribution by neighbor j, where $|\mathcal{N}_j|$ is the neighborhood cardinality of agent j. This essentially "averages" out the contributions of agent j so that one agent does not have more of an influence on the change in trust just because it has more neighbors.

At this point, as done in [21], we could also define a pairwise trust value, τ_{ij} and similarly define the trust dynamics for it in a straightforward manner. We focus on the neighborhood trust model in this work to avoid redundancy and refer the reader to [21] for further details.

IV. Invariance Result

If we define $\hat{\tau}$ to be the total trust in the system, that is

$$\hat{\tau} = 1^T\tau = \sum_{i=1}^{N} \tau_i, \tag{7}$$

then, similar to the self-centered trust modeling presented in [21], as a result of the coupled dynamics we get an invariance result relating the total trust to the performance of the system.

Theorem 1: Consider an N agent system with a static, connected interaction graph $\mathcal{G}$, where the agents' state dynamics are given by (4) and the trust dynamics are given by (6). Consider a performance cost $F(x)$ and the total trust $\hat{\tau}$ as defined in (3) and (7), respectively. Then, the following invariance holds

$$\frac{d}{dt}\left(F(x) + \hat{\tau}\right) = 0.$$

Proof: To prove this result let us notice that

$$\dot{\hat{\tau}} = \sum_{i=1}^{N} \dot{\tau}_i = -\sum_{i=1}^{N} \sum_{j\in\mathcal{N}_i} \frac{1}{|\mathcal{N}_j|} \frac{\partial F(x)}{\partial x_j}^T \dot{x}_j$$

where the term $\frac{1}{|\mathcal{N}_j|} \frac{\partial F(x)}{\partial x_j}^T \dot{x}_j$ has no dependence on i and appears in the summation once for every time that agent j is a neighbor of an agent i, or exactly $|\mathcal{N}_j|$ times. Consequently, the dynamics for $\hat{\tau}$ can be written as

$$\dot{\hat{\tau}} = -\sum_{i=1}^{N} \frac{\partial F(x)}{\partial x_i}^T \dot{x}_i. \tag{8}$$

Furthermore, if we write $F(x)$ from (3) in a slightly different manner, that is,

$$F(x) = \frac{1}{2} \sum_{i=1}^{N} \sum_{j\in\mathcal{N}_i} F_{ij}(\|x_i - x_j\|),$$

then the time derivative is given by

$$\frac{d}{dt}F(x) = \frac{1}{2} \sum_{i=1}^{N} \sum_{j\in\mathcal{N}_i} \left(\frac{\partial F_{ij}}{\partial x_i}^T \dot{x}_i + \frac{\partial F_{ij}}{\partial x_j}^T \dot{x}_j \right),$$

which, since the performance costs are symmetric, i.e. $F_{ij} = F_{ji}$, and the network is undirected, simplifies to

$$\frac{d}{dt}F(x) = \sum_{i=1}^{N}\sum_{j\in\mathcal{N}_i}\frac{\partial F_{ij}}{\partial x_i}^T \dot{x}_i. \tag{9}$$

and by substituting

$$\frac{\partial F(x)}{\partial x_i} = \sum_{j\in\mathcal{N}_i}\frac{\partial F_{ij}}{\partial x_i}$$

into the expression for $\dot{\hat{\tau}}$ in (8), we get $\frac{d}{dt}F(x) = -\dot{\hat{\tau}}$ and thus the desired result is obtained. ■

Therefore, in order to achieve the desired performance, i.e. minimize the cost $F(x)$, we must have the total trust in the system increase. In fact, the two are intimately linked, meaning that decreases in total trust correspond to increases in the cost, or decreased performance. This is consistent with the organizational psychology literature in which studies have shown a positive correlation between trust and performance in teams within organizations [18].

V. Convergence Properties

One of the results that was shown in [21] was that, for two agents executing rendezvous, we know under what specific conditions on the initial trust that the agents will converge and diverge. However, for the self-centered trust model, it was not possible to show general convergence guarantees. In fact, the two agent scenario is a special case in that it results in the self-centered trust model being identical to the team-oriented trust model because there is only one edge in the graph. With the new, team-oriented trust model, however, we can give explicit conditions under which the system will achieve the goal of minimizing the performance cost. In order to show the convergence results, we first discuss monotonicity properties of the trust values.

Lemma 1: Consider an N agent system where the agents' state dynamics are given in (4) and the trust dynamics are given in (6). Assume that the static interaction graph $\mathcal{G}$ is connected and that $\tau_i(0) > 0$, for all $i \in \{1,\dots,N\}$. Then τ_i is non-decreasing and $\tau_i(t) \geq \tau_i(0)$, for all $i \in \{1,\dots,N\}$ and for all $t > 0$.

Proof: We can rewrite (4) for agent j as

$$\dot{x}_j = -\tau_j\frac{\partial F(x)}{\partial x_j}$$

and substitute this into (5) to get the following expression for the trust dynamics,

$$\dot{\tau}_i = \sum_{j\in\mathcal{N}_i}\frac{\tau_j}{|\mathcal{N}_j|}\frac{\partial F(x)}{\partial x_j}^T\frac{\partial F(x)}{\partial x_j}$$

or, equivalently,

$$\dot{\tau}_i = \sum_{j\in\mathcal{N}_i}\frac{\tau_j}{|\mathcal{N}_j|}\left\|\frac{\partial F(x)}{\partial x_j}\right\|^2.$$

Since $\tau_j(0) > 0$, for all $j \in \{1,\dots,N\}$, $\dot{\tau}_i \geq 0$, for all $i \in \{1,\dots,N\}$, and it follows that τ_i is non-decreasing and $\tau_i(t) \geq \tau_i(0) > 0$, for all $i \in \{1,\dots,N\}$, for all $t > 0$. ■

This tells us that if the initial trust values are all positive, they will stay positive over the duration of the evolution of the dynamics and also finite due to the invariance in Theorem 1. Similarly, if all of the trust values are initially negative, they will remain negative, as shown in the following lemma.

Lemma 2: Consider an N agent system where the agents' state dynamics and trust dynamics are given in (4) and (6), respectively. Assume that the static interaction graph $\mathcal{G}$ is connected and $\tau_i(0) < 0$, for all $i \in \{1,\dots,N\}$. Then τ_i is non-increasing and $\tau_i(t) \leq \tau_i(0)$, for all $i \in \{1,\dots,N\}$ and for all $t > 0$.

Proof: This proof follows directly from the proof of Lemma 1 by simply flipping the inequalities. ■

We are now ready to discuss convergence results of the collaborative objective for the multi-agent system.

Theorem 2: Consider an N agent system where the agents' state dynamics are given in (4) and the trust dynamics are given in (6). Assume the underlying static interaction graph $\mathcal{G}$ is connected and $\tau_i(0) > 0$, for all $i \in \{1,\dots,N\}$. Consider a performance cost $F(x)$ and the total trust τ as defined in (3) and (7), respectively. Then $x(t)$ asymptotically converges to a local minimum of the performance cost $F(x)$.

Proof: The collective agent dynamics are given by

$$\dot{x}(t) = -B(\tau(t))\frac{\partial F(x(t))}{\partial x} \tag{10}$$

where

$$\frac{\partial F(x)}{\partial x} = \left[\frac{\partial F(x)}{\partial x_1}^T, \frac{\partial F(x)}{\partial x_2}^T, \dots, \frac{\partial F(x)}{\partial x_N}^T\right]^T$$

and $B(\tau(t)) = \mathrm{diag}(\tau(t)) \otimes I_d$ where $\mathrm{diag}(\tau(t)) \in \mathbb{R}^{N\times N}$ is a diagonal matrix with $\tau_i(t)$ as its i'th diagonal element, I_d is the $d \times d$ identity matrix, and $\otimes$ is the Kronecker product.

Taking the time derivative of the performance cost $F(x)$ yields

$$\frac{dF(x)}{dt} = \frac{\partial F(x)}{\partial x}^T\dot{x} = -\frac{\partial F(x)}{\partial x}^T B(\tau(t))\frac{\partial F(x)}{\partial x} \tag{11}$$

and, following the result from Lemma 1, $\tau_i(t) \geq \tau_i(0) > 0$, for all $i \in 1,\dots,N$. Therefore, $B(\tau(t))$ has all positive elements on the diagonal and thus is positive definite for all t, giving the desired result $\frac{dF}{dt} < 0$ for all $\left\|\frac{\partial F(x(t))}{\partial x}\right\| \neq 0$, and $\frac{dF}{dt} = 0$ if and only if $\left\|\frac{\partial F(x(t))}{\partial x}\right\| = 0$. At this point, since the agents are moving along the anti-gradient of the cost function and $\dot{x} = 0$ only when $\frac{\partial F(x)}{\partial x} = 0$, it follows that $x(t)$ converges to a local minimum of $F(x)$. ■

This result is straightforward in that the model was designed so that trust will increase as long as all agents are behaving correctly and as long as the trust values stay positive, the state dynamics are designed such that the agents will indeed behave correctly. We note that this result follows from the ideal, nominal case in which every agent

has perfect information about its neighbors and that all agents behave according to the state dynamics in (4). Future work will include cases where agents do not have perfect information about their neighbors' states and when there may be noisy measurements and possibly deceit.

The same type of analysis is done for a system where all of the trust values are initially negative, showing that the agents will not reach a configuration that corresponds to a local minimum of the cost function $F(x)$ and may actually cause the agents' state trajectories to diverge.

***Theorem** 3:* Consider an N agent system where the agents' state dynamics are given in (4) and the trust dynamics are given in (6). Assume the static interaction graph $\mathcal{G}$ is connected and $\tau_i(0) < 0$, for all $i \in 1, \ldots, N$. Consider a performance cost $F(x)$ and the total trust τ as defined in (3) and (7), respectively. Then the agents' states $x(t)$ will not converge to a local minimum of the performance cost $F(x)$.

Proof: Following Lemma 2, we know that $\tau_i(t) < 0$ for all $t \geq 0$ and all $i \in \{1, \ldots, N\}$. Hence, $B(\tau(t))$ in (10) is negative definite for all $t \geq 0$ and from (11), we know that $\frac{dF}{dt} > 0$ for all $\left\| \frac{\partial F(x)}{\partial x} \right\| \neq 0$. And, because every agent is updating its state in the direction corresponding to an increase in $F(x)$, two cases may arise according to the particular nature of the cost function $F(x)$ and the initial configuration of the system, that is either the agents will stop in a configuration corresponding to a local maximum for which $\left\| \frac{\partial F(x)}{\partial x} \right\| = 0$, or the agents' states will keep updating indefinitely while the performance cost $F(x)$ goes to infinity. Clearly, in either of these two cases the agents will not reach a local minimum of $F(x)$. ■

These theorems allow us to predict the behavior of the system when either all of the trust values are initially positive or all of them are initially negative. This suggests that trust systems should be initiated with positive trust in order to guarantee that the desired performance is achieved. Note that, if the initial trust values of the agents are mixed, i.e., both positive and negative, then following the previous analysis, we cannot make any claim about the definiteness of the matrix $B(\tau(t))$, and thus about the convergence properties of the system.

VI. Simulations

As a case study of the gradient-descent multi-agent framework discussed in this paper, we consider the seminal swarm aggregation work described in [26]. Swarm aggregation has been widely investigated by the robotics and control communities over the last two decades, e.g., [27], [28], and is a basic behavior that many swarms in nature exhibit, such as ant colonies, flocks of birds, and schools of fish. The reader is referred to [29] for a comprehensive overview of swarm stability and optimization.

As in [29], we consider the following performance cost,

$$F(x) = \sum_{i=1}^{N} \sum_{j \in \mathcal{N}_i} F_{ij}(x).$$

The pairwise performance cost $F_{ij}(x)$ is defined as

$$F_{ij}(x) = F_{\mathrm{a}}(\|x_i - x_j\|) - F_{\mathrm{r}}(\|x_i - x_j\|),$$

where $F_{\mathrm{a}}(\|x_i - x_j\|)$ and $F_{\mathrm{r}}(\|x_i - x_j\|)$ are the pairwise aggregate performance cost and repulsive performance cost, respectively. In particular, the following have been used

$$F_{\mathrm{a}}(\|x_i - x_j\|) = a \, \frac{\|x_i - x_j\|^2}{2}$$

and

$$F_{\mathrm{r}}(\|x_i - x_j\|) = b \, \log(\|x_i - x_j\|)$$

with a and b the aggregation and repulsion tuning parameters, respectively, which determines the size of the aggregation area (see again [29] for further details).

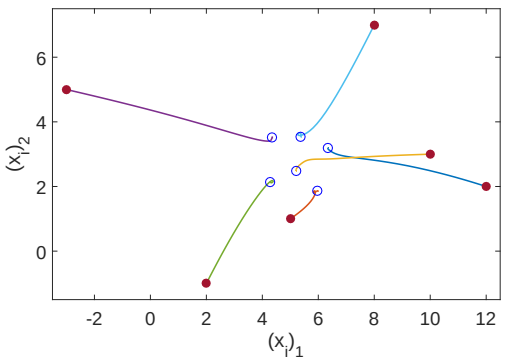
For purposes of the simulations in this section, we let the constants $a = 1$ and $b = 2$ and used 2-dimensional agents, i.e. $d = 2$. We first simulated the dynamics with $N = 6$ and all positive initial trust values. The interaction graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is defined such that $\mathcal{E} = \mathcal{E}_K \setminus \{(v_1, v_5), (v_2, v_4)\}$ where $\mathcal{E}_K$ is the edge set associated with a complete graph, $\mathcal{G}_K$. The initial stacked states are $x(0) = [12, 2, 5, 1, 10, 3, -3, 5, 2, -1, 8, 7]^T$ and $\tau(0) = [1, 2, 0.5, 2, 6, 0.2]^T$. Because all of the trust values are initially positive, this system converges to a minimum of $F(x)$, as proved in Theorem 2 and shown in Fig.2a.

We also simulated the dynamics with negative initial trust values in order to illustrate the negative result of Theorem 3. The same graph structure $\mathcal{G}$ and $x(0)$ as in the last example were used, but the trust values were set to be $\tau(0) = [-1, -2, -0.5, -2, -6, -0.2]^T$. In Fig. 2b, we show the evolution of the states for the first 20 time steps in the simulation, where you can see that the agents are moving away from each other, and they continue to do so for the times not pictured. In fact, the agents do not ever reach an equilibrium. Furthermore, the cost $F(x)$ goes to infinity over time and thus a local minimum is not reached.

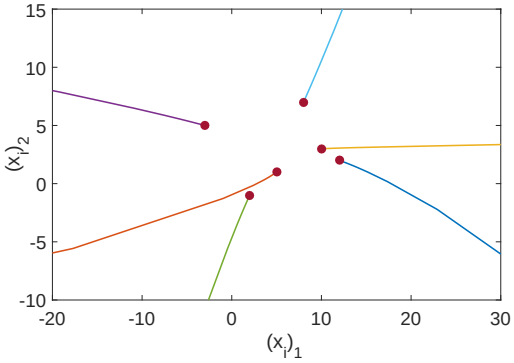
VII. Conclusions

In this work, we presented a trust model for multi-agent systems that takes into account each agent's overall contribution to achieving the team's goal. We showed that by coupling the trust dynamics to the state dynamics, an invariance results stating that an increase in performance must correspond to an increase in total trust in the system. We also showed that under certain initial trust conditions, the system is guaranteed to either have deteriorating performance or to achieve the team performance goal.

This trust model and the presented coupled dynamics are a stepping stone for many possible avenues for future work. One such avenue is to explore what happens when agents are malicious or misbehaving, as trust should be helpful in detecting and/or recovering from these types of situations. Another fruitful direction is to explore the effects of noise or corruption on the measurements made by an agent in calculating its change in trust. Each agent needs to obtain a value for much how its neighbors are contributing to the performance cost and this information may be noisy or



(a) All trust values initially positive.



(b) All trust values initially negative.

Fig. 2. Pictured are the state trajectories resulting from the dynamics for the trust-based algorithm, where both (a) and (b) use the same initial state (x) values, but different initial trust (τ) values. The filled circles represent the initial states and the empty circles, only in (a), represent the states at the end of the simulation. For (b), the agents continue to move away from one another and thus do not achieve the desired behavior.

may need to be estimated or communicated. Additionally, it would be insightful to develop topological conditions under which the team is guaranteed to achieve its performance goal regardless of the signs of the initial trust values. Lastly, we plan to explore heterogeneous dynamics in which robot agents do not operate with a trust metric, but human agents do, in order to capture the fact that humans and robots may have different behaviors.

REFERENCES

[1] J. Leitner, "Multi-robot cooperation in space: A survey," in *Advanced Technologies for Enhanced Quality of Life*, 2009.

[2] C. J. R. McCook and J. M. Esposito, "Flocking for heterogeneous robot swarms: A military convoy scenario," in *Southeastern Symposium on System Theory*, 2007.

[3] J. L. Baxter, E. K. Burke, J. M. Garibaldi, and M. Norman, *Multi-Robot Search and Rescue: A Potential Field Based Approach*. Springer Berlin Heidelberg, 2007, pp. 9–16.

[4] A. Tellaeche, I. Maurtua, and A. Ibarguren, "Human robot interaction in industrial robotics. examples from research centers to industry," in *IEEE Conference on Emerging Technologies Factory Automation (ETFA)*, 2015.

[5] C. D. Kidd and C. Breazeal, "Effect of a robot on user perceptions," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, vol. 4, 2004.

[6] L. Takayama and C. Pantofaru, "Influences on proxemic behaviors in human-robot interaction," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2009.

[7] A. Steinfeld, T. Fong, D. Kaber, M. Lewis, J. Scholtz, A. Schultz, and M. Goodrich, "Common metrics for human-robot interaction," in *ACM SIGCHI/SIGART Conference on Human-robot Interaction*, 2006.

[8] A. Freedy, E. DeVisser, G. Weltman, and N. Coeyman, "Measurement of trust in human-robot collaboration," in *International Symposium on Collaborative Technologies and Systems*, 2007.

[9] N. Wang, D. V. Pynadath, and S. G. Hill, "Trust calibration within a human-robot team: Comparing automatically generated explanations," in *ACM/IEEE International Conference on Human-Robot Interaction*, 2016.

[10] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. C. Chen, E. J. de Visser, and R. Parasuraman, "A meta-analysis of factors affecting trust in human-robot interaction," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 2011.

[11] A. Kolling, P. Walker, N. Chakraborty, K. Sycara, and M. Lewis, "Human interaction with robot swarms: A survey," *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 1, pp. 9–26, Feb 2016.

[12] J. Nagi, A. Giusti, L. M. Gambardella, and G. A. D. Caro, "Human-swarm interaction using spatial gestures," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2014.

[13] T. Setter, A. Fouraker, M. Egerstedt, and H. Kawashima, "Haptic interactions with multi-robot swarms using manipulability," *Journal of Human-Robot Interaction*, 2015.

[14] J. Y. C. Chen and M. J. Barnes, "Human-agent teaming for multirobot control: A review of human factors issues," *IEEE Transactions on Human-Machine Systems*, vol. 44, no. 1, pp. 13–29, Feb 2014.

[15] C. E. Harriott, A. E. Seiffert, S. T. Hayes, and J. A. Adams, "Biologically-inspired human-swarm interaction metrics," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 58, no. 1, pp. 1471–1475, 2014.

[16] X. Wang, Z. Shi, F. Zhang, and Y. Wang, "Mutual trust based scheduling for (semi)autonomous multi-agent systems," in *American Control Conference*, 2015.

[17] H. Saeidi, F. McLane, B. Sadrfaidpour, E. Sand, S. Fu, J. Rodriguez, J. R. Wagner, and Y. Wang, "Trust-based mixed-initiative teleoperation of mobile robots," in *2016 American Control Conference (ACC)*, July 2016, pp. 6177–6182.

[18] F. Erdem and J. Ozen, "Cognitive and affective dimensions of trust in developing team performance," *Team Performance Management: An International Journal*, vol. 9, pp. 131–135, 2003.

[19] K. S. Cook and O. Schilke, "The role of public, relational and organizational trust in economic affairs," *Corporate Reputation Review*, vol. 13, no. 2, pp. 98 – 109, 2010.

[20] L. J. Chang, B. B. Doll, M. van t Wout, M. J. Frank, and A. G. Sanfey, "Seeing is believing: Trustworthiness as a dynamic belief," *Cognitive Psychology*, vol. 61, no. 2, pp. 87–105, sep 2010.

[21] T. Setter, A. Gasparri, and M. Egerstedt, "Trust-based interactions in teams of mobile agents," in *American Control Conference*, 2016.

[22] R. Olfati-Saber and R. Murray, "Distributed structural stabilization and tracking for formations of dynamic multi-agents," in *IEEE Conf. Decision and Control*, 2002.

[23] M. Mesbahi and M. Egerstedt, *Graph Theoretic Methods in Multiagent Networks*. Princeton University Press, 2010.

[24] P. Ogren, M. Egerstedt, and X. Hu, "A control lyapunov function approach to multiagent coordination," *IEEE Trans. Robot. Autom.*, vol. 18, no. 5, pp. 847–851, Oct 2002.

[25] M. H. DeGroot, "Reaching a consensus," *Journal of the American Statistical Association*, vol. 69, p. 118121, 1974.

[26] V. Gazi and K. M. Passino, "Stability analysis of swarms," *IEEE Trans. on Automatic Control*, vol. 48, no. 4, pp. 692–697, 2003.

[27] A. Gasparri, G. Oriolo, A. Priolo, and G. Ulivi, "A swarm aggregation algorithm based on local interaction for multi-robot systems with actuator saturations," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012.

[28] W. Li and M. W. Spong, "Unified cooperative control of multiple agents on a sphere for different spherical patterns," *IEEE Trans. on Automatic Control*, vol. 59, no. 5, pp. 1283–1289, May 2014.

[29] V. Gazi and K. M. Passino, *Swarm Stability and Optimization*, 1st ed. Springer Publishing Co., Inc., 2011.