# Trust-Based Interactions in Teams of Mobile Agents

Tina Setter Andrea Gasparri Magnus Egerstedt

*Abstract*— **This paper introduces a trust model that couples the change in performance in a team of agents to how the agents perceive (or trust) each other. This combination of social dynamics and physical update laws not only changes the performance of the system, but it has the potential to make it deteriorate in a dramatic fashion. In fact, in the two-agent case, it is shown that the system exhibits finite escape time through an invariance result that carries over also to larger systems and more elaborate trust models. The invariance result states that an increase in performance must be accompanied by an increase in the total trust in the network (and vice versa for deteriorating performance). Finally, the connection is made between the proposed model and the belief and group polarization phenomena encountered in group processes driven by social interaction dynamics.**

## I. Introduction

Human-decision making processes are highly involved in that the group dynamics come into play in a sometimes surprising and counter-intuitive manner. For example, the phenomenon of *group polarization* [1] is well-established whereby members of a group collectively arrive at decisions and positions that are more extreme than any individually held position before the group convened. This effect has been observed in politics, radicalization of religious beliefs, and even jury deliberations, e.g., [2]. One implication from this phenomenon is that the group does not perform any direct averaging operation under these circumstances.

One caveat for group polarization to occur is that the group members share similar, initial views. If they instead start off far from each other, so-called *belief polarization* may occur [3], whereby the group dynamics make the team members distance themselves from each other over time. The classic study in this area investigated peoples' views regarding the death penalty [4]. This type of clustering of opinions is, for example, consistent with Krause's opinion dynamics model [5], [6], where peoples' opinions cluster based in part on how far away their initial opinions are from each other. (It should be noted though that Krause's model does not cause the decision makers' beliefs to get further away during the evolution of the process.)

These two aspects of human group dynamics have, by themselves, no real bearing on engineered multi-agent systems, where the interaction rules can be carefully designed to achieve the desired outcomes (See, e.g., [7], [8] and the references therein.) In fact, the field of distributed multi-agent systems can be roughly thought of as falling into one of two different camps, namely the design of distributed interaction laws for engineered systems, and the modeling of interaction dynamics in naturally occurring systems. In other words, multi-agent robotics concerns itself with the design of effective control laws for achieving coordinated objectives in a distributed manner, while opinion dynamics aims at understanding how opinions or preferences propagate through a network of individuals.

These two camps collide, however, when human decision makers are embedded within teams of robotic agents in that engineered interaction laws and human opinions will have to coexist. This may not be a major issue when a small number of highly trained operators are engaging with the robot team, but it has the potential to become more acute as the number of operators increases, or their skill level decreases. As a result, the study of human-swarm interactions has received significant attention during the last few years, e.g., [9], and a number of different interaction modalities have been proposed and studied, as summarized in [10]. Examples include leader-follower control, fluid-based manipulation, behavioral interactions, and boundary control. (See [11]–[15] for a representative sample.)

In this paper, we do not concern ourselves with the particular choice of interaction-abstraction. Instead, we investigate different aspects of what might happen when social opinion dynamics and multi-robot interaction laws coexist in a coupled manner. And, it should be noted already at this point that we do not claim to be accurate from a psychological modeling point-of-view. Instead, we are simply introducing a novel model that captures both the standard, engineered types of interaction rules in conjunction with the belief polarization aspects that may arise when people interact with teams of autonomous agents.

To illustrate the effect of the coupling between opinions and physical actions, we let autonomous agents' and humans' interaction laws be a function of how well the system is responding, or how much they *trust* the performance of the system. This question of trust is central to the study in this paper. And, although trust has been studied extensively, e.g., [16]–[18]; either as a design tool for mitigating effects of malicious behaviors or as a feature of the opinion dynamics, the coupling between group dynamics exhibiting belief polarization effects and multi-robot interaction laws in a dynamical systems manner is novel.

Rather than studying beliefs in isolation from the actions that they trigger, we couple the opinion dynamics with the physical dynamics through the trust, i.e., the trust evolution

The work by the first and third authors was sponsored by Grant No. N0014-15-1-2115 from the U.S. Office for Naval Research. Tina Setter and Magnus Egerstedt are with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA, {tsetter3,magnus}@gatech.edu. Andrea Gasparri is with the Engineering Department of the University of Rome "Roma Tre", Italy, gasparri@dia.uniroma3.it.

becomes an explicit function of how well neighboring agents are responding to individual actions. That is, agent $i$ will "trust" its neighbors more if they behave the way agent $i$ expects them to, which in turn changes how agent $i$ responds to its neighbors. As will be seen, this formulation will allow for belief and group polarization effects and has the potential to make the system performance deteriorate rapidly. In fact, if there is not sufficient initial trust between agents, the system may exhibit finite escape time.

The paper is organized as follows. In Section II, we provide a two-agent example as a cautionary tale of the complexity due to the coupling of social dynamics and physical update laws. In Section III, we introduce two general trust-based interaction models that couple trust evolution to how well adjacent agents are responding to an agent's movements. In Section IV, we demonstrate how the total trust in the network is linked to the performance of the system through an invariance result. In Section V, we draw final remarks.

## II. A Two-Agent Mood Picture

The reason why it is both problematic and worth-while to explicitly couple the agents' opinions about neighboring agents to their actions is that these types of coupled interaction effects (through both beliefs and physical states) will inevitably play some parts in future human-swarm interaction scenarios. But, these couplings are indeed quite delicate in that a lot of surprising effects can emerge. As a first illustration of this, consider the case of two agents with scalar states $x_i,\ i=1,2$, who are to meet at a joint location, i.e., by solving the, by now, classic rendezvous problem [19].

If we use the quantity $\frac{1}{2}(x_1-x_2)^2$ as a measure of how well (actually, how poorly) the two agents are doing, we note that the change in "performance" is given by

$$\frac{d}{dt}\left(\frac{1}{2}(x_1-x_2)^2\right)=(x_1-x_2)\dot{x}_1+(x_2-x_1)\dot{x}_2.$$

As the term $(x_1-x_2)\dot{x}_1$ encodes how much the movement of Agent 1 contributes to the change in the performance (and vice versa for Agent 2), the trust that Agent 2 "feels" towards Agent 1 should reflect this fact. In other words, if Agent 2 is contributing a lot to the agents getting closer, then Agent 1 should trust Agent 2 more.

A possible encoding of the previous observation could be to let the scalar trusts $\tau_1$ and $\tau_2$ evolve as

$$\dot{\tau}_1=-(x_2-x_1)\dot{x}_2,\quad \dot{\tau}_2=-(x_1-x_2)\dot{x}_1,$$

where the negative sign is used to describe the fact that a reduction in inter-agent distance should correspond to an increase in trust. Moreover, the trust itself needs to be coupled to the motion of the two agents. And, following the observation of group and belief polarization, we let a positive trust $\tau_1$ mean that the Agent 1 is indeed moving towards Agent 2, while a negative trust would mean the opposite. In the context of the consensus equation (e.g., [8]), this could be directly encoded using the trusts as weights, as done in the DeGroot Model in [20], thus yielding the composite system

$$\begin{array}{ll}\dot{x}_1=\tau_1(x_2-x_1) & \dot{x}_2=\tau_2(x_1-x_2)\\ \dot{\tau}_1=\tau_2(x_1-x_2)^2 & \dot{\tau}_2=\tau_1(x_1-x_2)^2,\end{array}\qquad(1)$$

which we note, due to the presence of square terms, is not globally Lipschitz, i.e., there might even be issues with the existence of solutions [21].

In fact, if we let $\xi=x_1-x_2$ denote the disagreement between the two agents, and $\hat{\tau}=\tau_1+\tau_2$ the total trust in the system, we note that

$$\begin{aligned}\dot{\xi}&=\dot{x}_1-\dot{x}_2=-(\tau_1+\tau_2)(x_1-x_2)=-\hat{\tau}\xi\\ \dot{\hat{\tau}}&=\dot{\tau}_1+\dot{\tau}_2=(\tau_1+\tau_2)(x_1-x_2)^2=\hat{\tau}\xi^2.\end{aligned}$$

From these equations we note that

$$\dot{\hat{\tau}}=-\xi\dot{\xi}\ \Rightarrow\ \hat{\tau}=-\frac{1}{2}\xi^2+c,$$

for some constant $c$. Therefore, $\hat{\tau}+1/2\xi^2$ is an invariant, a fact that will carry over to more complex models in subsequent sections, and we state this as a lemma.

***Lemma** 2.1:* Under the two-agent dynamics given in (1), the following holds

$$\frac{d}{dt}\left(\hat{\tau}+\frac{1}{2}\xi^2\right)=0.$$

We would like to understand when this system will result in the agents' states converging to the same point and when it will cause the states to diverge. Since $\xi(0)=0$ implies that an agreement is trivially reached initially and maintained throughout, we assume that $\xi(0)\neq 0$.

The first thing to note about this system is that if $\hat{\tau}>0$, then $\dot{\hat{\tau}}>0$ as long as $\xi\neq 0$. Thus, if $\hat{\tau}(0)>0$, the total trust in the system will increase monotonically. But the invariance $\xi^2=2(c-\hat{\tau})$ implies that this increase in $\hat{\tau}$ will have to correspond to a decrease in $\xi$. In other words, if $\hat{\tau}(0)>0$ then $\hat{\tau}$ will increase monotonically until, in the limit, $\xi=0$, i.e., the two agents will indeed agree asymptotically.

Similarly, if the initial, total trust is zero, then both $\dot{\hat{\tau}}=0$ and $\dot{\xi}=0$ and the disagreement does not change. (Even though both $x_1$ and $x_2$ do change at the same rate.) But what happens if the initial, total trust is less than zero?

Plugging the expression for $\hat{\tau}$ from Lemma 2.1 into $\dot{\xi}$ gives

$$\dot{\xi}=\frac{1}{2}\xi^3-c\xi.$$

Setting

$$\eta=\frac{1}{\xi^2}$$

yields

$$\dot{\eta}=-\frac{2}{\xi^3}\dot{\xi}=-1+\frac{2c}{\xi^2}=-1+2c\eta,$$

with solution (assuming $c\neq 0$)

$$\eta(t)=e^{2ct}\left(\eta(0)-\frac{1}{2c}\right)+\frac{1}{2c}.$$

We now note that $\eta(0)>0$ (as long as $\xi(0)\neq 0$) and two different cases must be investigated, namely when $c>0$ and when $c<0$. If $c<0$, $\eta(t)$ will decay exponentially from

$\eta(0) > 0$ to $1/(2c) < 0$, and thus cross $\eta = 0$ at some finite time. But, since $\eta = \xi^{-2}$, this implies that $\xi(t)$ exhibits finite escape time, i.e., it goes to $\pm\infty$ in finite time.

If $c > 0$ we note that $\hat{\tau}(0) < 0$ implies that

$$\hat{\tau}(0) = -\frac{1}{2}\xi^2(0)+c < 0 \;\Rightarrow\; \eta(0) < \frac{1}{2c} \;\Rightarrow\; \eta(0)-\frac{1}{2c} < 0.$$

So in this case, $\eta$ starts at $\eta(0) > 0$ and then decays exponentially to $-\infty$. As such, also in this case, there exists a finite time at which $\eta = 0$, i.e., also in this case does $\xi(t)$ exhibit finite escape time. (As a final note, if $c = 0$ then $\eta(t) = -t + \eta(0)$, i.e., at time $t = \eta(0)$, the error dynamics escapes to $\pm\infty$.)

We have thus established the following two-agent theorem:

***Theorem** 2.1:* Consider the two-agent system

$$\begin{array}{ll} \dot{x}_1 = \tau_1(x_2 - x_1) & \dot{x}_2 = \tau_2(x_1 - x_2) \\ \dot{\tau}_1 = \tau_2(x_1 - x_2)^2 & \dot{\tau}_2 = \tau_1(x_1 - x_2)^2. \end{array}$$

If $\tau_1(0) + \tau_2(0) > 0$ then $\lim_{t\to\infty} |x_1(t) - x_2(t)| = 0$. If $\tau_1(0) + \tau_2(0) = 0$ then $x_1(t) - x_2(t)$ is constant. Finally, if $\tau_1(0) + \tau_2(0) < 0$ then $|x_1(t) - x_2(t)|$ diverges to infinity in finite time, whenever $x_1(0) \neq x_2(0)$.

These phenomena are shown in Figure 1, where two cases are shown; one where $\hat{\tau}(0) > 0$, and the agent agree asymptotically, and one where $\hat{\tau}(0) < 0$, and they diverge in finite time.

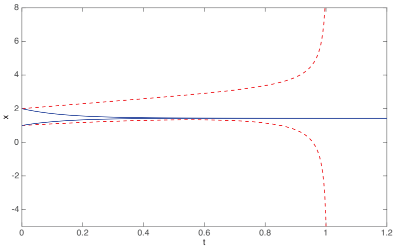


Fig. 1. Two different scenarios are shown. The first (middle trajectories) corresponds to the total, initial trust being positive, causing the two agents to reach an agreement asymptotically. The second case corresponds to a negative initial trust, resulting in diverging states in finite time.

Through the addition of an innocent-looking trust dynamics coupled to the update laws for the agent states, not only can the system diverge, it may diverge in finite time! And, returning to the issue of belief polarization, this means that if two agents do not trust each other sufficiently much initially, the process deteriorates completely. If, for example, one were to add a cut-off, as is done in Krause's model, i.e., the agents only take each other into account if their distrust is not too great, polarized states are achieved that are moreover more "extreme" than the agents' initial states, as shown in Fig. 2. In other words, belief polarization results.

## III. Coupled Trust Models

Following the development in the previous section, one can now define a more general trust-based interaction model that couples the trust evolution to how well adjacent agents are responding to an agent's movements. By specifying the desired network performance through a more general, pairwise, symmetric, inter-agent performance cost $F_{ij}(\|x_i - x_j\|)$, as is done, for example, in the formation control literature, e.g., [22]–[24], the corresponding contribution by Agent $j$ to the increase in cost is given by

$$\frac{\partial F_{ij}(\|x_i - x_j\|)}{\partial x_j}\dot{x}_j,$$

which, in turn, should be coupled to the trust evolution. (Note that the states need no longer be scalar.)

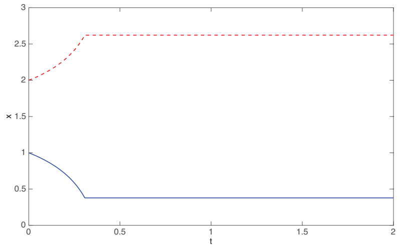


Fig. 2. Belief polarization is achieved with the agents assuming more extreme positions than their initial positions by only paying attention to each other if the trust values are above a certain threshold.

There are different ways in which this expression can be coupled to a trust model. For example, human trust is typically pairwise, i.e., a person does not necessarily trust all people equally. As such, a study of human-to-human-interactions must capture this pairwise relationship. However, what makes human-to-autonomous-agent-interactions different is that the trust is more uniform, i.e., a person may or may not trust the autonomous agents but will not necessarily be able to tell agents apart or form pairwise opinions about the performance of the agents. As such, we need two different types of models that reflect these two different types of trusts.

### A. Pairwise Trust

Consider a collection of $N$ agents, interacting over a static, undirected, and connected information-exchange network, $G = (V, E)$, where the vertex set $V = \{1, \ldots, N\}$ is the set of agents and the edge set $E \subset V \times V$ is a set of unordered pairs that encode the adjacency in the network. Each agent has a physical state $x_i$, $i = 1, \ldots, N$, and we add an additional state $\tau_{ij}$ to each ordered agent-pair in the network, which denotes Agent $i$'s level of trust for an adjacent Agent $j$. And, in light of the previous discussion, given a pairwise performance cost, $F_{ij}(\|x_i - x_j\|)$, we let the evolution of $\tau_{ij}$ depend on how much Agent $j$'s movement makes the performance cost decrease, i.e.,

$$\dot{\tau}_{ij} = -\frac{\partial F_{ij}(\|x_i - x_j\|)}{\partial x_j}\dot{x}_j,$$

or, in the case of the rendezvous problem, with

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

we get

$$\dot{\tau}_{ij} = (x_i - x_j)^T\dot{x}_j.$$

### B. Neighborhood Trust

Under pairwise trust, i.e., how much Agent $i$ trusts Agent $j$, the number of states could potentially grow very large as the network grows. Moreover, in a network of largely anonymous agents, pairwise relationships are, as already discussed, not a realistic feature. For these reasons, we can instead let trust be a neighborhood property, i.e., how much Agent $i$ trusts its neighbors.

Following the previous construction, we let $\tau_i$ denote Agent $i$'s trust level, and use an aggregated update law

$$\dot{\tau}_i = -\sum_{j \in N_i} \frac{\partial F_{ij}(\|x_i - x_j\|)}{\partial x_j} \dot{x}_j,$$

where $N_i$ is the set of agents adjacent to Agent $i$ in the network., i.e., $N_i = \{j \in V \mid (i,j) \in E\}$. In the case of the rendezvous problem, this simplifies to

$$\dot{\tau}_i = \sum_{j \in N_i} (x_i - x_j)^T \dot{x}_j.$$

### C. Connecting Trust to State Evolution

These two trust models, both of which will be considered, must now be coupled to the evolution of the physical states. To this end, we note that in the absence of any trust states, a standard, gradient-descent-based update law is given by

$$\dot{x}_i = -\sum_{j \in N_i} \frac{\partial F_{ij}(\|x_i - x_j\|)}{\partial x_i},$$

which has been employed repeatedly in the literature for a number of different types of applications, including formation control, connectivity maintenance, and collision-avoidance, e.g., [22]. However, in this paper, we chose to augment this model by adding a trust gain to the evolution, i.e.,

$$\dot{x}_i = -\sum_{j \in N_i} \tau_{ij} \frac{\partial F_{ij}(\|x_i - x_j\|)}{\partial x_i},$$

or

$$\dot{x}_i = -\tau_i \sum_{j \in N_i} \frac{\partial F_{ij}(\|x_i - x_j\|)}{\partial x_i},$$

depending on which trust model we use. For the rendezvous problem, these system equations become

$$\dot{x}_i = \sum_{j \in N_i} \tau_{ij}(x_j - x_i),$$

or

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

### D. Belief Polarization

If one wants to capture the belief polarization phenomenon, then these update laws simply have to be adjusted to ensure that only neighboring agents that are sufficiently trusted are taken into account. If we let $\bar{\tau}$ denote this threshold (typically a negative number), the update equations for the neighborhood trust scenario become

$$\dot{x}_i = \begin{cases} -\tau_i \sum_{j \in N_i} \frac{\partial F_{ij}(\|x_i - x_j\|)}{\partial x_i} & \text{if } \tau_i \geq \bar{\tau} \\ 0 & \text{otherwise.} \end{cases}$$

For pairwise trust, we redefine the neighborhood as

$$N_i(\tau) = \{j \in V \mid (i,j) \in E \text{ and } \tau_{ij} \geq \bar{\tau}\},$$

with the update law becoming

$$\dot{x}_i = -\sum_{j \in N_i(\tau)} \tau_{ij} \frac{\partial F_{ij}(\|x_i - x_j\|)}{\partial x_i}. \tag{2}$$

Examples of running (2) over 5 agents are shown in Fig. 3, with random initial conditions over $x$ (uniform over $[0,1]$) and $\tau$ (normally distributed with zero mean), with $\bar{\tau} = 0.3$. As shown, depending on the initial states and trusts, dramatically different results are obtained.

## IV. Invariance Results

One common feature of the trust model variations previously discussed is that the total trust in the network is intimately linked to the performance of the system through an invariance, as shown for $N = 2$. We first present such a result for the rendezvous problem under the collective trust model and then we generalize it to the case for which the network performance is evaluated through a desired performance cost.

### A. Consensus With Collective Trust

Consider a system composed of $N$ agents solving the rendezvous problem under the collective trust model:

$$\begin{aligned} \dot{x}_i &= \tau_i \sum_{j \in N_i} (x_j - x_i) \\ \dot{\tau}_i &= \sum_{j \in N_i} \left[ \tau_j (x_j - x_i)^T \left( \sum_{k \in N_j} (x_j - x_k) \right) \right], \end{aligned} \tag{3}$$

where we have substituted the expressions for $\dot{x}_j$ in the trust model. Let us assume that the individual states are scalars[1], and let us set $x = [x_1, \ldots, x_N]^T$ and $\tau = [\tau_1, \ldots, \tau_N]^T$. For this system, the following invariance result holds

***Lemma 4.1:*** Consider a collection of $N$ agents under the dynamics in (3), with $x_i \in \mathbb{R}$, then the following holds

$$\frac{d}{dt}\left( \frac{1}{2} \|D^T x\|^2 + \mathbf{1}^T \tau \right) = 0.$$

with $D$ the incidence matrix obtained by associating an arbitrary orientation with the network topology.

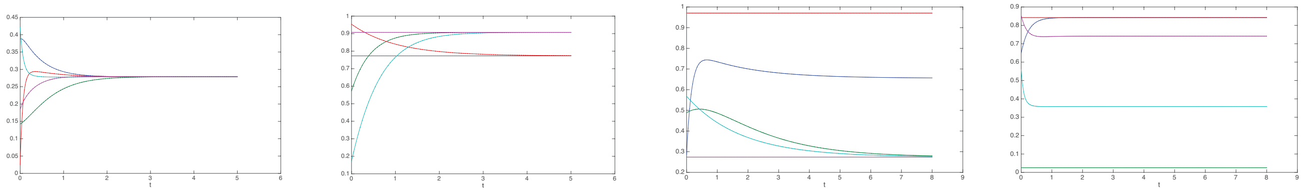
*Proof:* Let us define the total trust $\hat{\tau}$ in the network as

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

and let us notice that the evolution of the total trust $\hat{\tau}$ is

$$\dot{\hat{\tau}} = \sum_{i=1}^{N} \left\{ \sum_{j \in N_i} \left[ \tau_j (x_j - x_i)^T \left( \sum_{k \in N_j} (x_j - x_k) \right) \right] \right\},$$

[1] Note that, this assumption is by no means a restriction. Indeed, all it does is make the notation less complex



(a) (b) (c) (d)

Fig. 3. Four different outcomes are shown, where five agents cluster into one, two, three, and even four different groups, respectively, as a result of the random initial conditions on the state and trust values.

which, by rearranging the summation order (since the network is undirected), can be rewritten as

$$\begin{aligned}\dot{\hat{\tau}} &= \sum_{j=1}^{N} \tau_j \left[ \sum_{i\in N_j} (x_j - x_i)^T \left( \sum_{k\in N_j} (x_j - x_k) \right) \right] \\ &= \sum_{j=1}^{N} \left[ \tau_j \Big\| \sum_{i\in N_j} (x_j - x_i) \Big\|^2 \right].\end{aligned}$$

Now, letting $L$ be the Laplacian associated with the information-exchange network, we have that

$$\mathbf{1}^T \dot{\tau} = x^T L \mathcal{T} L x,$$

where $\mathbf{1} = [1, \ldots, 1]^T$ and $\mathcal{T} = \text{diag}(\tau)$. We moreover observe that the $x$-dynamics becomes

$$\dot{x} = -\mathcal{T} L x.$$

Furthermore, note that

$$\frac{1}{2}\frac{d}{dt}\|D^T x\|^2 = \frac{1}{2}\frac{d}{dt}(x^T D D^T x) = x^T L \dot{x} = -x^T L \mathcal{T} L x,$$

which is exactly equal to $-\mathbf{1}^T\dot{\tau}$, and thus the result follows. ■

We point out that this invariance result is directly analogous to the two-agent invariance result given in Lemma 2.1, and it tells us that an overall increase in performance must correspond to an increase in the total trust in the network. Notably, it turns out that this holds true also in the more general cases, whereby the network performance is evaluated through the performance cost

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

as discussed in the following two subsections.

### *B. The General Neighborhood Trust Case*

Let us consider the neighborhood-based trust model, for which the coupled dynamics of an agent $i$ is

$$\begin{aligned}\dot{x}_i &= -\tau_i \sum_{j\in N_i} \frac{\partial F_{ij}(\|x_i - x_j\|)}{\partial x_i}, \\ \dot{\tau}_i &= -\sum_{j\in N_i} \frac{\partial F_{ij}(\|x_i - x_j\|)}{\partial x_j}\dot{x}_j.\end{aligned} \tag{4}$$

Then, the following invariance result holds.

***Theorem*** *4.1:* Consider a collection of $N$ agents under the dynamics in (4), with $x_i \in \mathbb{R}$, then the following holds

$$\frac{d}{dt}\big(F(x) + \mathbf{1}^T\tau\big) = 0.$$

*Proof:* To prove this result, let us notice that the derivative of the performance cost $F(x(t))$ is

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

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

$$\frac{d}{dt}F(x(t)) = \sum_{i=1}^{N}\sum_{j\in N_i}\frac{\partial F_{ij}}{\partial x_j}\dot{x}_j. \tag{5}$$

Similarly, the total trust evolution is given by

$$\dot{\hat{\tau}} = \sum_{i=1}^{N}\dot{\tau}_i = -\sum_{i=1}^{N}\sum_{j\in N_i}\frac{\partial F_{ij}}{\partial x_j}\dot{x}_j,$$

Therefore, the result follows. ■

### *C. The General Pairwise Trust Case*

Let us consider the pairwise-based trust model, for which the coupled dynamics of an agent $i$ is

$$\begin{aligned}\dot{x}_i &= -\sum_{j\in N_i} \tau_{ij}\frac{\partial F_{ij}(\|x_i - x_j\|)}{\partial x_i}, \\ \dot{\tau}_{ij} &= -\frac{\partial F_{ij}(\|x_i - x_j\|)}{\partial x_j}\dot{x}_j,\end{aligned} \tag{6}$$

and for which the total trust in the network is redefined as

$$\hat{\tau} = \sum_{i=1}^{N}\sum_{j\in N_i}\tau_{ij}.$$

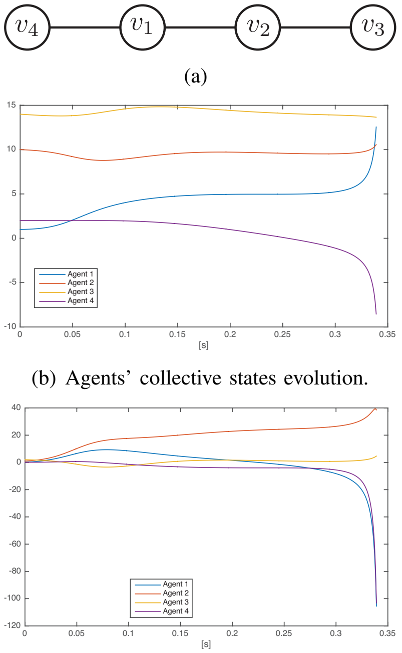
Then, the following invariance result holds

***Theorem*** *4.2:* Consider a collection of $N$ agents under the dynamics in (6), with $x_i \in \mathbb{R}$, then the following holds

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

*Proof:* To prove this result, let us notice that the total trust evolution is given by

$$\dot{\hat{\tau}} = \sum_{i=1}^{N}\sum_{j\in N_i}\dot{\tau}_{ij} = -\sum_{i=1}^{N}\sum_{j\in N_i}\frac{\partial F_{ij}}{\partial x_j}\dot{x}_j.$$



(a)

(b) Agents' collective states evolution.

(c) Agents' trust states evolution

Fig. 4. Simulation involving four agents interacting according to a line-based graph topology. It can be noticed how a finite-escape time occurs at time 0.35s.

and by combining that with (5), the result follows. ■

### D. Interpretations

The invariance results show that the performance of the network is intimately tied to the evolution of the total trust. If there is not sufficient trust between agents, the opinions may diverge. Furthermore, the general $N$-agents case, with $N >$2, is much more complex than the two-agents case.

For example, consider a system of four agents for which the interactions are dictated by the line topology depicted in Fig. 4(a). Assume that the agents are to solve the rendezvous problem under the collective trust model as in (3) and consider the following set of initial conditions

$$x(0) = \left[1\ 10\ 14\ 2\right]^T, \quad \tau(0) = \left[.1\ 1\ 2\ .1\right]^T. \tag{7}$$

Figs. 4(b) and 4(c) depict the collective dynamics and the trust dynamics over time, respectively. It can be noticed that a finite-escape time occurs around $t = .35s$ even though the initial trusts of the agents are positive and so is their sum.

Theorem 2.1 only provides a result concerning the evolution of the sum of the agents' trusts. Notably, for the two-agent case with consensus dynamics, this also suffices to constrain the evolution of the two agents' trusts. Unfortunately, this bind between the evolution of the sum and the evolution of the agents' trusts no longer exists for the $N > 2$, and we cannot prevent a finite-escape time to occur simply by looking at the initial sum of trusts.

## V. Conclusion

In this paper we introduced the idea of trust-based interactions to model the dynamics of a system where humans and autonomous agents coexist and interact. We proposed two different trust models, i.e., pairwise and neighborhood, to describe different scenarios, e.g., human-to-human and human-to-swarm interactions. We showed how the total trust in the network is intimately linked to the performance of the system through an invariance and we also demonstrated through illustrative examples that such a combination of social dynamics and physical update laws not only changes the performance of the system, but it has the potential to make the performance deteriorate in a dramatic fashion.

## References

[1] E. Aronson, *Social Psychology*. Prentice Hall, 2010.

[2] D. Isenberg, "Group polarization: A critical review and meta-analysis," *J. Pers. Soc. Psychol.*, vol. 50, pp. 1141–1151, 1986.

[3] T. Kelly, "Disagreement, dogmatism, and belief polarization," *Journal of Philosophy*, vol. 105, pp. 611–633, 2008.

[4] C. Lord, L. Ross, and M. Lepper, "Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence," *J. Pers. Soc. Psychol.*, vol. 37, pp. 2098–2109, 1979.

[5] V. Blondel, J. Hendrickx, and J. Tsitsiklis, "On krause's multi-agent consensus model with state-dependent connectivity," *IEEE Trans. Autom. Control*, vol. 54, pp. 2586–2597, 2009.

[6] U. Krause, "A discrete nonlinear and non-autonomous model of consensus formation," in *Proc. Commun. Difference Equations*. Gordon and Breach Pub, 2000, pp. 227–236.

[7] F. Bullo, J. Cortes, and S. Martinez, *Distributed Control of Robotic Networks*. Princeton University Press, 2009.

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

[9] M. Hagele, W. Schaaf, and E. Helms, "Robot assistants at manual workplaces: Effective co-operation and safety aspects," in *Proceedings of the 33rd International Symposium on Robotics*, 2002.

[10] J. de la Croix and M. Egerstedt, "A control lyapunov function approach to human-swarm interactions," in *American Control Conference*, 2015.

[11] M. Egerstedt *et al.*, *Large-Scale Networks in Engineering and Life Sciences*. Birkhauser, 2015, ch. Interacting with Networks of Mobile Agents, pp. 199–224.

[12] Z. Kira and M. Potter, "Exerting human control over decentralized robot swarms," in *Autonomous Robots and Agents*, 2009.

[13] A. Kolling *et al.*, "Human swarm interaction: An experimental study of two types of interaction with foraging swarms," *Journal of Human-Robot Interaction*, vol. 2, pp. 103–128, 2013.

[14] J. McLurkin *et al.*, "Speaking swarmish:human-robot interface design for large swarms of autonomous mobile robots," in *AAAI Spring Symp.*, 2006.

[15] G. Podevijn, M. Dorigo, and M. Dorigo, "Self-organised feedback in human swarm interaction," in *Workshop on Robot Feedback in Human-Robot Interaction*, 2012.

[16] X. Liu and J. Baras, "Using trust in distributed consensus with adversaries in sensor and other networks," in *Information Fusion, 17th International Conference on*, July 2014, pp. 1–7.

[17] G. Theodorakopoulos and J. Baras, "On trust models and trust evaluation metrics for ad hoc networks," *Journal on Selected Areas in Communications*, vol. 24, no. 2, pp. 318–328, Feb 2006.

[18] Y. Wang and J. Vassileva, "Bayesian network-based trust model," in *Web Intelligence. IEEE/WIC International Conf. on*, 2003.

[19] H. Ando *et al.*, "Distributed memoryless point convergence algorithm for mobile robots with limited visibility," *IEEE Trans. Robot. Autom.*, vol. 15, pp. 818–828, 1999.

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

[21] H. K. Khalil, *Nonlinear Systems*. Prentice Hall, 2002.

[22] M. Ji and M. Egerstedt, "Distributed coordination control of multiagent systems while preserving connectedness," *IEEE Trans. on Robotics*, vol. 23, no. 4, pp. 693–703, 2007.

[23] 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.

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