

This One Simple Trick Disrupts Digital Communities

Philip Feldman

ASRC Federal

Columbia, USA

philip.feldman@asrcfederal.com

Aaron Dant

ASRC Federal

Columbia, USA

aaron.dant@asrcfederal.com

Wayne Lutters

Dept. of Human Centered Computing

University of MD, Baltimore County

Catonsville, USA

lutters@umbc.edu

Abstract—This paper describes an agent based simulation used to model human actions in belief space, a high-dimensional subset of information space associated with opinions. Using insights from animal collective behavior, we are able to simulate and identify behavior patterns that are similar to nomadic, flocking and stampeding patterns of animal groups. These behaviors have analogous manifestations in human interaction, emerging as solitary explorers, the fashion-conscious, and echo chambers, whose members are only aware of each other. We demonstrate that a small portion of nomadic agents that widely traverse belief space can disrupt a larger population of stampeding agents. We then model the concept of Adversarial Herding, where trolls, adversaries or other bad actors can exploit properties of technologically mediated communication to artificially create self sustaining runaway polarization. We call this condition the *Pishkin Effect* as it recalls the large scale buffalo stampedes that could be created by native Americans hunters. We then discuss opportunities for system design that could leverage the ability to recognize these negative patterns, and discuss affordances that may disrupt the formation of natural and deliberate echo chambers.

Index Terms—belief space, computer simulation, echo chamber, flocking, group processes, polarization, social behavior

I. INTRODUCTION

Emergent collective behavior such as flocking, schooling and swarming is a curiously universal phenomena. In addition to animal behavior, social influence has been studied in such diverse contexts as coupled oscillators, consensus formation in networks, load balancing, and belief propagation [1]. Collective behavior based on social influence clearly has benefits. Being in a group offers protection from predators and can be helpful in finding food, particularly in cases where traces are faint [2]. Human collective actions extends these behaviors into information domains, where they can manifest in the writing of government constitutions [3], revolutions [4], as well as more trivial, but socially important examples such as fashion [5].

Collective behavior exists on a continuum. At one end, nomadic patterns have individuals or small groups dispersed over large areas. Though these nomads do not enjoy the protection of large groups, the overall population gains resilience due to its extended footprint. Further along the spectrum are various types of clustering - flocks, herds and schools. This higher level of social interaction can increase overall information processing leading to better sensing and more optimal ex-

ploitation of resources [2]. At the far end of the spectrum, social influence becomes the dominant factor, outweighing such things as environmental cues. These conditions lead to social inertia, where the collective is unable to adapt to changing conditions. In the extreme, this can lead to panics and stampedes [6]. Often, to balance out nomadic costs and the risks of social inertia, a population will incorporate an explore/exploit strategy [7], where a small portion of the population will range beyond the traditional habitat. This strategy provides a level of resilience for the overall population while letting the majority exploit opportunities discovered by the nomads.

Stampedes by animals, a mass movement in a unified direction [8] is analogous to echo chambers or groupthink, a form of extreme cohesion that creates its own social reality [9]. The most dramatic case of this may be Nazi Germany, where Arendt wrote of “[a] movement which is being propelled with increasing speed in a certain direction” [10]. Moscovici and Doise [11] describe how groups in free discussion compile complex information into simplified norms as a mechanism to achieve agreement and polarization. Norms are the direction that the group aligns with, and have a relationship to the group’s coherence and rate of change. Munson and Resnik [12] show online user behavior (confirming, diversity-seeking, diversity-avoiding) that provides a computer-mediated framework these user actions.

When information is mediated through technology, the information horizons that inform where to “explore” and when to “exploit” can become obscured, leading to destructive behaviors. Technology makes available abundant information with wildly varying levels of veracity, structure, bias, and credibility, blurring the lines between naive belief and well-supported evidence. Beliefs are not facts, but they are *factive* – they *feel* like facts [13]. This makes prima facie determination of the quality of specific information extremely difficult for humans and systems that depend on them [14]. Like stampedes in the animal kingdom, echo chambers, the self-reinforcing reflection of a belief [15], are a symptom of this misalignment of environmental awareness and social influence. In many cases, credible-looking, low-quality or misleading information contributes to the formation of echo chambers [16]. Members often believe that they have access to all needed information and are unaware of critical perspectives [17].

Information retrieval (IR) systems often sidestep the issue of information quality by tailoring results to user preferences, thus trusting in users' abilities to discern fact from fiction or opinion [18]. Yet human assessments are vulnerable to cognitive biases, social identities, perceived pressure, norms, and other factors that interfere with accurate judgment [19]. The process of seeking information can lead to a self-confirmatory feedback loop, in which low quality information is perceived as valid and higher quality information is excluded until it closes off into an echo chamber. Once sealed in an internal discourse, the echo chamber can gain velocity and become a stampede. Our information technologies provide many tools to aid this process. Framed in a supporting context with a believable interface, large groups of people can be persuaded of many things - that a president is secretly from another country [20] or that aliens have landed in New Jersey [21].

The ability for IR-mediated group interactions to connect like-minded people who can't easily see out of their filter bubbles can lead to events such as the "Pizzagate conspiracy", an echo chamber based on group belief that there was high ranking Democrat involvement in human trafficking [22]. This echo chamber produced extremely dangerous real world outcomes [23]. Some solutions to this issue have been tried successfully. Chandrasekharan et al have shown that banning Reddit hate groups can disrupt echo chambers with continuing positive effects on post-membership users as they are re-integrated into a larger, more diverse community [24]. Salganik has shown that ranking that prioritizes quality over popularity [25] can eliminate runaway effects in music selection. It seems reasonable that design changes at social websites like Facebook, Twitter or Reddit, to enhance information diversity [26] could decrease the likelihood of such a runaway result.

To take all these separate pieces and understand how they function together, we need a model that can represent, in a simplified fashion, critical aspects of computer mediated group interaction. For this, we turn to agent-based simulation. Agent-based and cellular automata-based simulation has proven to be a particularly effective mechanism for modeling the complex interplay between individuals in a population. In problems ranging from neighborhood segregation [27] to opinion dynamics [28] to culture dissemination [29], these types of simulations have been shown to be effective in generating complex emergent collective behaviors from sets of simple, understandable rules applied to the agent.

II. PREVIOUS WORK

Animal models have often served as a starting point for understanding human interaction with information. Danchen et. al. showed that animals and humans both use inadvertent social information (ISI) to influence decisions about environmental quality and appropriateness [30]. Card and Pirolli [31] successfully demonstrated the utility of animal models for individual human information foraging behaviors. Deneubourg and Goss' [32] work related to animal group cognitive behaviors such as flocks and herds. More recently, Olfati-Saber et. al. have shown that social influence leading to collective behaviors is

a widespread phenomenon in natural and artificial systems [1]. Connecting animal models to technology-mediated human group interaction, Belz et. al. have shown the emergence of spontaneous flocking in computer mediated communication [33].

The study of human group behavior has roots in the 19th-century work of LeBon [34], who showed that crowds can move and think like single organisms, which was later studied experimentally by Moscovici [11]. More recently, Krause [28] and Bikhchandani [35] have modeled opinion dynamics and echo chambers while Salganik [25] has demonstrated that online rating based on popularity can produce runaway results. Epstein et. al shows similar results for Information retrieval with the Search Engine Manipulation Effect [36].

Game theory has also explored this problem space, particularly with respect to the evolution of cooperation. Using the *Iterated Prisoner's Dilemma*, research by Nowak and others show that there is dominant transition pattern that establishes from an initial random population. The first population is the antisocial *Always Defect* (AllD), which transitions to more social *Tit-for-Tat* (TFT), and continues through *generous TFT* to the highly social, efficient, and vulnerable *Always Cooperate* which in turn can be decimated by AllD [37]. A more stable strategy is *win-stay, lose-shift* [38], where profitable strategies are maintained until they fail, at which point the agent explores other tactics. This resonates with aspects of the *Multi-Armed Bandit Problem*, which examines when an agent should exploit a current slot machine, or explore for a better one [39]. Bacharach [40] explored a theoretical explanations for the above behavior which he extended to coordination games such as Stag Hunt. His insight was that when humans identify with groups, they can reason from that perspective and rationally choose options such as cooperate.

Lastly, Curran [5] shows qualities such as fads in fashion (a form of flocking behavior) can also be represented in a spatial way using movement in a belief space by observing norms for skirt length and width. She shows a detailed example of collective cognitive movement through a belief space traced over 36 years.

These models are effective and compelling descriptions, but are not optimal for producing emergent patterns. Bonabeau [41] states that, "By definition, [emergent phenomena] cannot be reduced to the system's parts: the whole is more than the sum of its parts because of the interactions between the parts." Reynolds [42], Cucker [43], Olfati-Saber [44] and others have built and described agent-based simulations that produce emergent flocking, schooling, and herding characteristics that closely mimic observed animal behavior.

III. MODEL CONSIDERATIONS

Our base model explores two ideas: 1) that human navigation through *belief space* (a subset of information space associated with opinions) is analogous to animal motion through physical space, and 2) that the *digital inadvertent social information* (DISI) provided by humans interacting

with the belief environment can be used to characterize the underlying space.

We rest this assumption on recent work on neural coupling [45] which has been used to show that social cognition is a *physical* process where individual minds synchronize neural firing patterns to support mutual cognition [46]. Since Olfati-Saber et. al. have shown that the underlying mathematics of collective behavior are shared across a wide range of constrained and unconstrained consensus domains [1], we believe that mechanisms for navigating physical space can be extended to handle cognitive spaces. After extensive model development, we determined the minimum set of features that an agent requires re:

- 1) *Dimension* – The number of beliefs that a person may hold is not limited by physical space. They may hold opinions on many subjects. To keep calculations manageable, we set an upper bound of 10 dimensions.
- 2) *Velocity* – Humans and animals dynamically interact with their physical and information environments. Although they may have regions or territories that they prefer, movement in the physical and political sense is a defining characteristic.
- 3) *Heading* – There appears to be a rate-limited alignment component that is needed for a group to coalesce. This is obvious in the physical patterns of flocking or schooling, but also manifests in language (e.g. “political alignment”) [13] and fashion [5].
- 4) *Influence* – Agents within a specified range are capable of influencing each other’s orientation and speed, inversely proportional to distance. This in turn influences heading, as more aligned agents have more time to influence each other [1].

There is considerable work in using groups of agents to evaluate fitness landscapes, where agents behave as particles, and increase their attractiveness based on their height in the landscape [47]. This works well for difficult machine learning problems such as hyperparameter tuning [48], but it does not capture the aspect of alignment in community behavior that is observed in sociological contexts. For our purposes, we chose the Reynolds model [42], which uses velocity and heading alignment as major components of its flocking and herding model. We modified the algorithm for high-dimensional belief spaces as we discuss in the methods section.

With respect to the individual agents, we can adjust what we term the *social influence horizon* (SIH), or the area that the agent considers in its calculations. Closer agents have more influence over current agent’s orientation and velocity. A low radius means that the agent has less social influence, which encourages exploration of the environment. The larger the radius becomes, the more the agent is dominated by social influence at the expense of environmental considerations.

The environment that these agents operate in represents the belief space that is mediated by technology. It supports variable, asymmetric visibility of one agent to another, boundary characteristics and the overall size and number of

dimensions. In addition, it supports the storage of agent data at n-dimensional coordinates in the space.

A. Adversarial Herding

Geils [49] notes how social media can be used to increase polarization based on *emergent* poles. In other words, “normal” opposing views can be amplified by attentive bad actors, with the goal of causing generalized disruption. We added capabilities to the simulation that amplify the influence of selected individual agents to evaluate the effectiveness of potential disruptive mechanisms. Real world examples of this activity have been documented in the news media, where accounts of Russian troll farms organizing opposing groups [50]. Most recently, Stella et al. have uncovered examples of bot-augmented human actors in the 2018 Italian general election [51].

Though the term “Information Warfare” has been used in this context, we believe that a more appropriate metaphor is *adversarial herding*. War describes conflict between two or more internally organized entities. Herding, on the other hand, can be defined as one entity causing many unwilling or unwitting entities to move in a desired direction [52]. As with animals, herding can be performed directly, as with traditional methods using trained dogs to control livestock, or herding can be performed using subterfuge, as when native Americans would lure and drive buffalo over *Pishkin cliffs* [53]. We believe that this “herding” approach more resembles what is known about social network exploitation. As such, we extended our base model to observe how that might manifest in a social media environment where it is easy to hide one’s true identity [54]

Based on information collected from news stories and Gerasimov’s work on nonmilitary methods in conflict development [55], we added the ability to enable herding in the simulation, based on the following rules:

- *Herders can amplify arbitrary agents*, since they are not emotionally invested in their belief space position or orientation
- *Herders appear like multiple individuals* (sockpuppets or sybils) that may seem close and trustworthy, but they are actually a distant monolithic entity that is aware of a much larger belief space.
- *Herders amplify arbitrary pre-existing positions*. The insight is that *they are not herding in a direction, but to increase polarization*

IV. METHODS

This study consisted of data generation and subsequent data analysis. We built a stand-alone simulator using Java that created the multidimensional belief environment and then populated it with agents. All agents are double buffered so that there are no sequential artifacts from agent interaction. The program can be run in interactive or batch mode. Sampled output was saved to Excel files.

A. Simulation

The main components to be implemented were the agents and the environment they exist in. Since the number of beliefs that a person may hold is not limited by physical space, arbitrary numbers of dimensions need to be accommodated. This was accomplished by collecting one-dimensional *statements* into a structure defined as a *belief*. Agents move through space based on Reynold’s boids model [42], but with collision terms removed since individuals can hold identical views.

Each statement resembles a single element of an opinion dynamics model such as the ones used by Krause [28]. For this work, each dimension was considered equivalent. Though this model is naive in that it is linear and orthogonal, social distance interactions across dissimilar spaces have been examined by Bogunia [56] and Schwammle [57]. They show that our approach can be extended to handle non-linear spaces.

B. Position and orientation in high-dimensional belief spaces

The agents’ position vector of statements is updated by the orientation multiplied by the elapsed time since the last update (Equation 1):

$$\vec{a}c\vec{p} = \vec{a}\vec{p}p + (\vec{a}\hat{p})(vp)(dt) \quad (1)$$

Where $\vec{a}\hat{p}$ is the previous orientation unit vector, $\vec{a}c\vec{p}$ is the current position vector, $\vec{a}\vec{p}p$ is the previous position vector, dt is elapsed time, and vp is the previous velocity of the agent.

The agent’s target “orientation” vector of statements is updated by taking the average orientation of all agents that are within the SIH of the agent. This radius can be set to an arbitrary value. The influence drops off linearly until the radius is reached (Equation 2 - notation from [58]). The distinct phases identified in the results are exclusively the result of manipulating this radius.

$$\vec{t}\hat{o}_x = \frac{\sum_{n=1}^{n=\max[n \neq x]} \vec{a}\hat{p}_n (1 - \frac{\|w_x \vec{a}\vec{p}p_x - w_n \vec{a}\vec{p}p_n\|}{r})}{1 - \sum_{n=1}^{n=\max[n \neq x]} \|w_x \vec{a}\vec{p}p_n - w_n \vec{a}\vec{p}p_x\|} \quad (2)$$

$$\|w_x \vec{a}\vec{p}p_x - w_n \vec{a}\vec{p}p_n\| < w_n r$$

Where $\vec{t}\hat{o}$ is the target orientation vector, $\vec{a}\hat{p}$ is the previous orientation unit vector, $\vec{a}\vec{p}p$ is the previous position, and r is the social influence horizon. The w term is the scalar weight of the agent’s influence, which is set to 1.0 in non-herding cases.

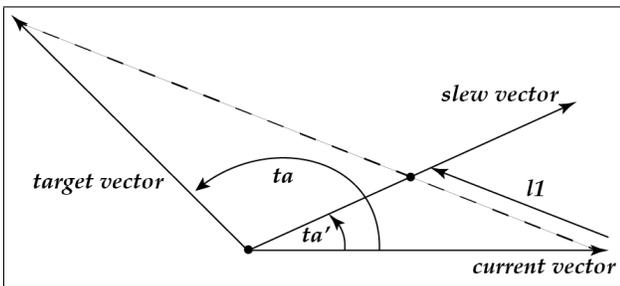


Fig. 1. Calculation of projected “slew” vector

The agent then interpolates its orientation towards the goal vector as a function of time (figure 1). First, the angle between the current and target orientation is calculated as a direction cosine. This allows for 2D calculations in the plane of rotation (Equation 3):

$$ta = \cos^{-1} \left(\frac{\vec{a}\hat{p} \cdot \vec{t}\hat{o}}{\|\vec{t}\hat{o}\|} \right) \quad (3)$$

Where ta is the 2D angle, $\vec{a}\hat{p}$ is the previous orientation unit vector, and $\vec{t}\hat{o}$ is the target orientation vector

The maximum incremental angle that the agent can rotate through (ta') is calculated as a fraction of the above, constrained by the turn rate of the agent and elapsed time since the last simulation step (Equation 4):

$$ta' = (ta)(rate)(dt) \quad (4)$$

The line formed by the current vector and the target vector can then be intersected with the slew vector. This intersection allows us to calculate the length of the vector that we add to the current n -dimensional orientation vector to produce the new orientation. Lastly, the scaled vector is added to the agent vector and normalized.

This model produces distinct emergent patterns, by adjusting only the SIH::

Nomadic Phase (figure 2) - A low SIH means low influence by other agents, so each agent moves in their own direction engaging in the “explore” phase of the Multi-Armed Bandit problem.

Flocking Phase (figure 3) - An intermediate SIH results in an agent whose movement is affected by nearby individuals. There is alignment with neighbors, but there is sufficient diversity so that orientations change over time

Stampede Phase (figure 4) - At high SIH, all members are exposed equally to each other in what is essentially a fully connected graph, which synchronize easily [1]. As such, alignment can become total and supports runaway [59] conditions. In the Multi-Armed Bandit problem this would be “exploit” without any explore. This is also represented in other literature as “filter bubbles”, “echo chambers” [17], “group polarization”, and “extremism” [11].

To support experimentation in this simulated domain, the simulation was configured so that the following variables could be manipulated.

- Number of agents
- One or two populations, with separate social influence horizons (0 - 10 units), operating independently or aware of each other
- Data accumulation in the environment by n -dimensional cell. Heatmaps, etc.
- Number of dimensions (2 - 10)
- Environment size (0 - 10 units)
- Environment border:
 - NONE: No effect when reaching the limit of the environment

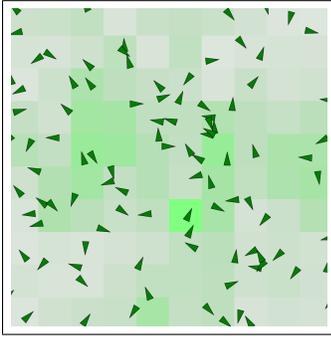


Fig. 2. Nomadic Phase

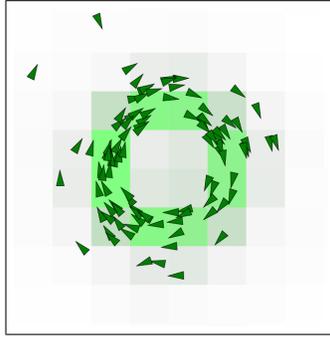


Fig. 3. Flocking Phase

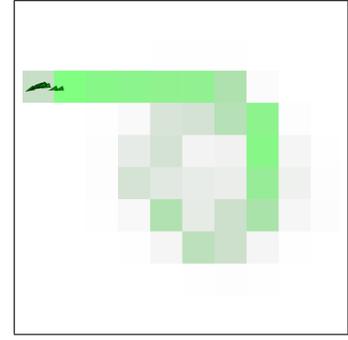


Fig. 4. Stampede Phase

- REFLECT: Agents are reflected back whenever they cross a border
- RESPAWN: If a border is crossed, the agent is re-initialized in the environment with a random heading, position and speed.

C. Adversarial Herding

To create inputs that resemble hypothetical herding behaviors, and to be able to view them, the model was extended to accommodate the following additional capabilities:

- The selected agent’s weight and social influence are increased to w' and r' . w' represents amplification by trolls, bots, etc. A large r' means that the bots can swamp other, normally “closer” signals. This models the effect of a monolithic entity controlling thousands of bots across the belief space [51].
- There are three optional herding modes:
 - Where the agent closest to the average heading is amplified. This mimics the effect of bots retweeting human actors that align with the adversaries’ goals.
 - A random “leader” is chosen for the duration of the simulation. This mimics the effect of sustained support of a single individual over time by bots and sybils.
 - A randomly chosen agent is amplified each cycle. This does not mimic any known technique, but was chosen because it is a simple case that randomly contaminates the social influence radius that is used to calculate group interaction.

D. Finding Patterns

Though the DISI patterns described above were visible and apparent when using the system interactively, the goal is to be able to detect the patterns programmatically. Our starting point is the work of Boguna et.al [56], who developed a set of models based on a mathematical abstraction of “social distance”. We used Dynamic Time Warping (DTW) [60] to determine the overall social distance between paths of individual agents. DTW works by determining how to transform one sequence into optimal alignment with another using non-linear mapping. This is different from linear transformations such as Least Squares, which produce an approximate fit. As outputs, DTW

produces a “warping path” and a total “warping distance”. This approach was sufficient to discriminate between the populations, while being fast enough to be used for large-scale analytics.

V. RESULTS

All experiments were managed by configuration files for repeatability, with typical experiments consisting of 100 agents run 10 times through each set of conditions. Initial experiments were done to determine if the number of dimensions altered agent behaviors. We found that the SIH had to be multiplied by the square root of the number of dimensions to produce the same agent behaviors. This is an example of the “curse of dimensionality” [61], which refers to the difficulty of calculating meaningful Euclidean distance in high-dimensional space. This may explain why polarization only happens after concepts have been simplified. Open discussion can be a form of dimension reduction [11] which creates a common framing that supports consensus [40]. Based on this result, the majority of simulations were run in two dimensions which eliminates projection artifacts from the visualizations.

Regardless of the number of dimensions (2 - 10 tested), we were able to see that agent behavior rapidly manifested in three phases by varying only the SIH. These three phases can be seen in Figures 2, 3 and 4.

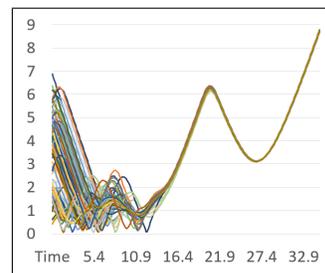


Fig. 5. Stampede distance

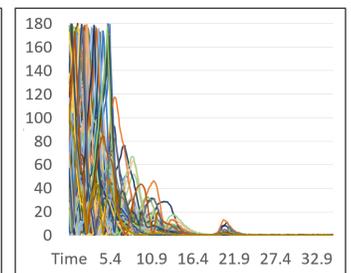


Fig. 6. Stampede heading

For these simulation runs, agents were initialized on a range of (-5.0, 5.0) on each dimension. A reflective barrier was placed at (-10.0, 10.0). This embodies the intuition that many concepts have inherent limits. For example, in fashion, a skirt has practical limits in length and width [5].

The first phase is determined entirely by the initial random generation of the agents. They continue along their paths until they encounter the containing barrier. The behavior is random with no emergent pattern. The second phase is the richest, characterized by the emergence of *flocks* or *schools*. The third phase represents an example of a runaway polarization condition or *stampede*. Figure 4 shows a tightly clustered group heading towards the edge of the environment. Figures 5 shows the convergence of the agents as they cluster with respect to the belief space origin. Figure 6 shows the same event with respect to heading. All agents become tightly aligned and clustered, and their position in space becomes more extreme over time. The only thing preventing the polarized group from heading off into infinity is the boundary.

A. Emergent Group Behavior

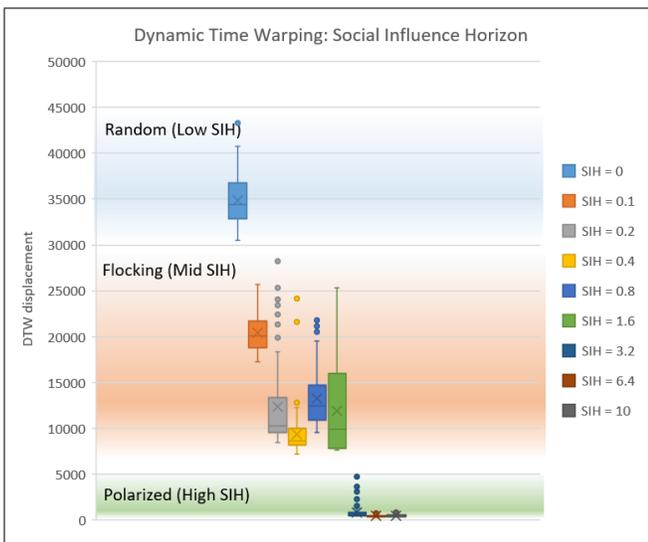


Fig. 7. Nomad, Flocking and Stampede DTW

We used Dynamic Time Warping (DTW) as discussed in the methods section and implemented in Java-ML to determine population membership with respect to SIH. DTW attempts to find the lowest distance that one set of points need to be moved to exactly match another sequence of points. We then build a matrix for the DTW distance between each agent and sum each column to compute the overall distance of one agent to the other agents in the simulation. The distribution of DTW distance by agent SIH is shown in Figure 7. The phase changes (nomad, flocking, stampede) are distinctive and *non-overlapping* in our datasets.

This strongly supports the observation that there are three distinct phases of behavior in the agents that are affected by the size of the SIH. Further, these phases are sufficiently distinct that they can be statistically recognized.

B. Interactions between populations

We also examined the interactions between two populations, each with a different SIH. Multiple studies across different disciplines ranging from neurology [62] to computer-human

interaction [12] have shown that populations often have explorer and exploiter subgroups. In nature, many effective strategies revolve around a majority exploit/minority explore pattern [62].

In one set of simulations, 10% of the population were given zero SIH, letting them explore the environment uninfluenced, while the other 90% were given the highest SIH (the size of the environment), which in prior runs had resulted in the group polarization of figure 4. These percentages reflect the results found by Cohen [62] as well as the percentage of diverse news consumers found by Flaxman et. al. in their study of browser logs [17].

The results of mixing these populations was startling. Although still tightly clustered, the stampede group would rarely encounter the simulation boundary and would instead be influenced towards the center by the presence of nomads. Although these agents were still a polarized group, they were no longer in a runaway condition (Figure 8).

The reason that the *nomads* are so successful in adjusting the trajectory of polarized groups has to do with their distribution. Because nomads maintain their random orientations throughout the simulation, they effectively provide a normal probability distribution over the environment within each dimension. What this model implies is that a sufficiently diverse population covers an belief space in such a way that if their *position* is visible to another population, it can have the effect of providing an attraction to the center of the environment.

C. Adversarial Herding

Adding arbitrary amplification to a single “super agent” as a way of creating a Pishkin-style stampede appears to be very effective. The impact of herding on DTW measures is shown in figure 11. The “No Herding” populations are the same populations as shown in figure 7 at the low end of the “flocking” SIH (the green middle band). The middle “Partial Herding” depicts the herding algorithm on the 0.1 unit influence horizon. From left to right, the herding implementation is:

- 1) Single, randomly chosen agent
- 2) The agent currently closest to the average flock heading
- 3) A random agent chosen at every sample

The least effective strategy is to choose a single agent. Somewhat counterintuitively, the most effective polarizing strategy is to choose random agents and amplify them for a short period. This between-strategies relationship is maintained in the green “Effective Herding” region, though the overall clustering is tighter and the variance between the results is much lower.

The reason for this is that randomly iterating over the entire population gives the entire population a high “effective” SIH. In the example shown in figure 7, the average DTW for the maximum SIH population evaluated is approximately half of the lowest population in figure 11. This seems reasonable, since the “natural” case is where all the agents have identical horizons should converge faster than any process that depends on a single agent, no matter how effectively distributed.

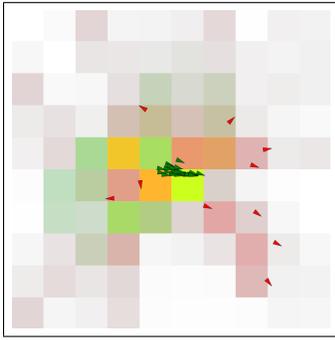


Fig. 8. Nomad influence



Fig. 9. Pishkin Effect

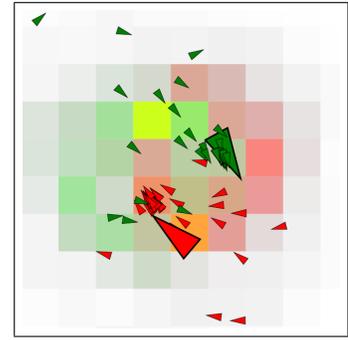


Fig. 10. Opposing Herding

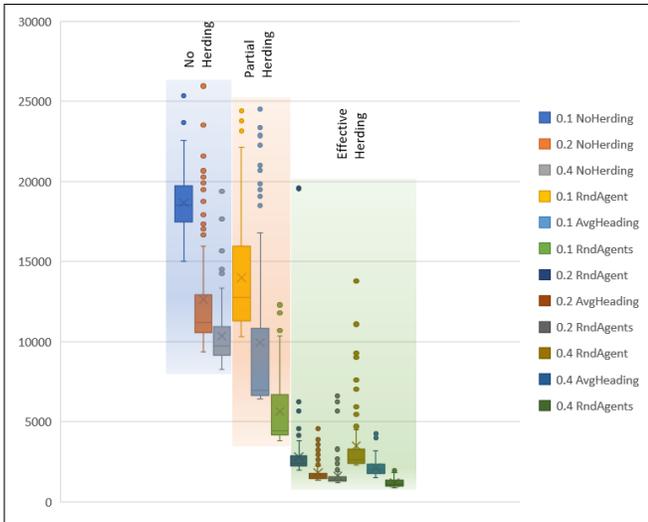


Fig. 11. No, Partial, and Effective Herding DTW

When a single, random agent is chosen, the stampede effect ends soon after the super agent encounters a lethal condition (The RESPAWN boundary in these simulations). Once free of the agent’s amplified influence, the other agents settle back into their normal flocking behavior. In the case where random agents are amplified, the results are more complicated. In the case where direction is being most influenced by completely random agents, the tendency towards polarization is maintained in the population, though it is unfocused, and takes a while to build back to the runaway condition. In the case where the super agent is chosen that most matches the average heading of the group, a more stable condition can arise where the clustering, orientation and velocity of the stampede is maintained independently of the particular agents involved. As stampeding agents encounter the lethal condition, they are continuously replaced by new agents that are randomly initialized into the environment to replace the terminated ones, but under the influence of the current super agent, these new agents are influenced by the ongoing stampede pattern and head back to the “cliff”.

To see how opposing polarized groups might be created, we also implemented a variant of herding implementation(2),

where an agent in a separate population that most matches the inverse vector of the first population is amplified (figure 10). This mimics the effect of an adversary supporting antagonistic extremism between groups. This resembles the RU-IRA interferences with the #blacklivesmatter/#bluelivesmatter twitter interaction [50].

VI. DISCUSSION

We have built a framework, based on a Reynolds [42] agent-based simulation to explore the emergent group behaviors based on interactions between individuals’ heading and orientation in a belief space hypercube. This simulation is a starting place to generate identifiable behaviors in belief space. We have begun work to look for similar patterns in the data produced by human users in subsequent studies. As our understanding of navigation in belief space improves, the model will be updated with more sophisticated rules grounded in observed human behavior patterns.

Animals make decisions as groups that manifest as a variety of patterns, including swarms, flocks, schools, and stampedes. The inadvertent social information provided by these groups allow other animals to infer ecosystem qualities. Humans appear to traverse belief space analogous to how animals move in physical space. Heading, velocity, and influence distance are qualities that we find are important to this process and occur in a wide variety of contexts. Recent support for this intuition have come from studies of group neural alignment by Stephens [45], Gallotti [46] and others. In simulation, the patterns of *digital inadvertent social information* (DISI) can be detected efficiently and on large data sets using DTW. Since analyzing DISI does not depend on language-dependent text analytics, this approach should be fundamentally domain independent.

The number of dimensions seems to matter. Low dimensions make it easier to initiate a stampede. Moscovici and Diose showed that unconstrained group deliberation will eventually reduce the subject of the discussion to a simplified representation, within which the group can polarize around a consensus. When structures are added to restrict this simplification process, group deliberation often leads to compromise, rather than polarization [11].

Individuals and groups must make decisions with incomplete information. Using social cues as a proxy for direct

evaluation of a problem has the benefit of less computation and shared risk, but contains the threat of groupthink. How this interaction plays out affects individual and group behavior in a dynamic, emergent way. Biological systems appear to have evolved the explore/exploit behavior pattern to address this issue. In a population, some relatively small percentage is biased towards nomadic exploration, while the majority tend to exploit their situation. This makes evolutionary sense, as exploration is risky. However, if a catastrophe strikes the main population, the species can rebuild from the nomad diaspora.

Stampedes are highly dynamic but fundamentally simple systems where individual environmental interaction is overridden by a single social reality [63]. Once established stampedes often continue until they encounter a sufficient disrupting condition. At a societal level in humans, stampedes, or echo chambers can be an existential threat to ongoing governmental and cultural norms [10], [64]. Increased awareness of a small population of nomadic explorers can positively influence the stable equilibrium of a stampede.

VII. IMPLICATIONS FOR DESIGN

We are influenced by psychological rules we can't control. For example, the title of this paper plays with our need to know "hidden knowledge". Similarly, the gaps that appear as we apply our understanding of physical spaces to virtual domains are inherently dangerous and exploitable.

The simple trick is this: *Changing the awareness of others with different levels of alignment disrupts communities*. This can be positive or negative. It can drive a polarizing group towards extremism, or hinder a stampeding mob.

The reason that this trick works is that when technology mediates communication, it changes the social cues that let us determine the trustworthiness of the information we receive. When we transfer our understanding of trustworthiness from the richness of the physical world to the sparse abstraction of online environments, the boundaries between well-supported evidence and naive belief become obscured. These low-dimensional spaces can be breeding grounds for misinformation.

As we have seen with the Jade Helm conspiracy theory [65], these "stampedes" can be naturally occurring. However, they can also be accentuated by an adversary that simply *amplifies* credible voices in the community, increasing the reach of the misinformation and bringing the credulous into alignment.

Poor information at an individual level that leads to Echo chambers are a problem. Echo chambers at a societal level can be an existential threat to ongoing governmental and cultural norms. Computer-mediated communication is very much a personalized interaction, but with effects that scale to populations. What changes could be made to these individual interactions to have desired socio-organizational and cultural effects?

Our research has indicated that an awareness of nomadic/explorer activity in belief space may help nudge stampeding groups away from a terminal trajectory and back towards "average" beliefs. Tajfel [66] states that groups can

exist "in opposition", so providing counter-narratives may be ineffective. Rather, we think that a potential approach to reducing online polarization is to inject diversity into users social media, search results, and video feeds [67]. The infrastructure exists for this already in platform's support of advertising. An existing template could be the Public Service Announcement (PSA).

US Broadcasters since 1927, have been obligated to "serve the public interest" in exchange for spectrum rights. One way that this has been addressed is through the creation of the PSA, "the purpose of which is to improve the health, safety, welfare, or enhancement of peoples lives and the more effective and beneficial functioning of their community, state or region" [68].

We believe that PSAs can be repurposed to support diversity injection (DI) through the following:

- 1) Random, non-political content designed to expand information horizons, analogous to clicking the "random article" link on Wikipedia.
- 2) Progressive levels of detail starting with an informative "hook" presented in social feeds or search results. Users should be able to explore as much or little as they want.
- 3) Simultaneous presentation to large populations. Google has been approximating this with their "doodle" since 1998, with widespread positive feedback, which indicates that there may be good receptivity to common serendipitous information.
- 4) Format should reflect the medium, i.e. text, images and videos.
- 5) Content should be easily verifiable, recognizable, and difficult to spoof.

We believe that such DI mechanisms as described above can serve as a "first do no harm" initial step in addressing the current crisis of misinformation. By nudging users towards an increased awareness of a wider world, DI interferes with the processes that lead to belief stampedes by increasing the number of dimensions, and awareness of different paths that others are taking.

As we gain deeper understanding of the mechanisms that influence group behaviors, it may be possible to further refine our designs and interfaces so that they no longer promote extremism at the socio-organizational level while still providing value at individual and small group levels. Generally, designs need to take into account how individual interaction manifests at different social scales. We believe that many of the current issues that are plaguing the largest-scale information providers, such as Facebook and Google derive from a failure to design for large scale belief behavior, as implemented in software and business models.

VIII. CONCLUSIONS

The benefits of cooperative structures appear continuously in evolution. Examples abound from cellular organelles to multicellular organisms to flocks, schools, forests and even ecosystems [69]. In these natural systems, mechanisms ranging from scent, to sight to sound evolved to provide relevant and

appropriate information for the collective to make optimal decisions [70]. An example of this is a school of fish able to find food when the scent is faint and patchy [2]. Though generally effective, collective approaches can fail spectacularly. Even with co-evolved information and social systems, buffalo stampede, whales beach, and locusts swarm into plagues.

Emergent structures for human social evolution have also developed, from tribe through village to city, nation-state and pan-national movement. As our social structures increased in size and sophistication, technology has provided us tools with an ever-increasing power to manipulate, store and transfer information. But our biological mechanisms for dealing with what we experience through our screens hasn't changed much since our ancestors were making tools from flint and obsidian. Approaches that worked when we were members of small tribes hunting and gathering can easily be fooled when the sources of our information are obscured, either by technological and economic constraints or with more malevolent intent. For cooperative structures to function, the quality of information provided by the members of the group must be judged with respect to the source's trustworthiness, not just its credibility. When the information flow between agents is distorted or misrepresented, "belief stampedes" can result. Examples include religious cults, market bubbles and crashes, and political phenomena like totalitarianism.

In addition to animal and human behavior, it could be wise to consider how these widespread and generalizable relationships might affect intelligent machines. The "flash crash" of 2010 could be regarded as a stampede of trading programs [71]. Given the rise of robotic systems such as self driving cars, it could be wise to consider the ramifications of these tendencies in populations of autonomous vehicles.

Recently, it has become popular to revisit moral dilemmas such as the Trolley Problem [72] in the light of progress with intelligent machines [73] [74]. Does a self-driving car decide to crash, risking its occupant to save the life of a child that it sees in its path? Although the authors certainly believe that this thinking is valuable, we also feel it is important to think about such problems when they occur at scale. Where does the moral responsibility lie? With the individual machine? Or the emergent, collective intelligence of the interlinked system?

Runaway group behavior can emerge if the conditions are right. Just as a canyon reduces the options for a herd of buffalo, making it easier for them to be spooked into stampeding, so too a reduction in the options available to intelligent systems can have profound effects. In the massive fires that swept through greater Los Angeles in December of 2017, police determined that GPS systems were attempting to re-route users around the closed highway 405 through burning neighborhoods [75]. It doesn't take much imagination to extrapolate what could happen if instead of human operators, autonomous vehicles with no training on how to behave in such massive disasters were the primary form of transportation. Without a pre-trained awareness of the dangers of flames and heat, large numbers of vehicles would obediently follow their carefully optimized route into the fire. As they enter the inferno, their

sensors begin to malfunction. The antennas that transmit position burn off and the cars become invisible to the network so the path continues to classify as clear and open. Vehicles by the hundreds continue to flow towards this virtual Pishkin, as betrayed by their design as buffalo were betrayed by their instincts hundreds of years before.

REFERENCES

- [1] R. Olfati-Saber, J. A. Fax, and R. M. Murray, "Consensus and cooperation in networked multi-agent systems," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 215–233, 2007.
- [2] D. Grünbaum, "Schooling as a strategy for taxis in a noisy environment," *Evolutionary Ecology*, vol. 12, no. 5, pp. 503–522, 1998.
- [3] A. Rutherford, Y. Lupu, M. Cebrian, I. Rahwan, B. LeVeck, and M. Garcia-Herranz, "Disentangling network and global effects in constitutional political development," *arXiv preprint arXiv:1606.04012*, 2016.
- [4] M. Lynch, D. Freelon, and S. Aday, "Online clustering, fear and uncertainty in egypt's transition," *Democratization*, pp. 1–19, 2017.
- [5] L. Curran, "An analysis of cycles in skirt lengths and widths in the uk and germany, 1954-1990," *Clothing and Textiles Research Journal*, vol. 17, no. 2, pp. 65–72, 1999.
- [6] C. J. Torney, T. Lorenzi, I. D. Couzin, and S. A. Levin, "Social information use and the evolution of unresponsiveness in collective systems," *Journal of the Royal Society Interface*, vol. 12, no. 103, p. 20140893, 2015.
- [7] O. Berger-Tal, J. Nathan, E. Meron, and D. Saltz, "The exploration-exploitation dilemma: a multidisciplinary framework," *PLoS one*, vol. 9, no. 4, p. e95693, 2014.
- [8] F. E. Ward, *The cowboy at work: All about his job and how he does it*. University of Oklahoma Press, 1987.
- [9] J. L. Burnette, J. M. Pollack, and D. R. Forsyth, "Leadership in extreme contexts: A groupthink analysis of the may 1996 mount everest disaster," *Journal of Leadership Studies*, vol. 4, no. 4, pp. 29–40, 2011.
- [10] H. Arendt, *The origins of totalitarianism*. Houghton Mifflin Harcourt, 1973, vol. 244.
- [11] S. Moscovici and W. Doise, *Conflict and consensus: A general theory of collective decisions*. Sage, 1994.
- [12] S. A. Munson and P. Resnick, "Presenting diverse political opinions: how and how much," in *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 2010, pp. 1457–1466.
- [13] D. R. DeNicola, *Understanding Ignorance: The Surprising Impact of What We Don't Know*. MIT Press, 2017.
- [14] T. Lukoianova and V. L. Rubin, "Veracity roadmap: Is big data objective, truthful and credible?" 2014.
- [15] V. O. Key, *The responsible electorate*, 3rd ed. Belknap Press of Harvard University Press, 1968.
- [16] N. JafariNaimi and E. M. Meyers, "Collective intelligence or group think?: Engaging participation patterns in world without oil," in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 2015, pp. 1872–1881.
- [17] S. Flaxman, S. Goel, and J. M. Rao, "Filter bubbles, echo chambers, and online news consumption," *Public Opinion Quarterly*, vol. 80, no. S1, pp. 298–320, 2016.
- [18] A. C. Nied, L. Stewart, E. Spiro, and K. Starbird, "Alternative narratives of crisis events: Communities and social botnets engaged on social media," in *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 2017, pp. 263–266.
- [19] R. Ahmad and W. G. Lutters, "Interpreting user-generated content: what makes a blog believable?" in *HCI LNCS*, vol. 6778. Springer, 2011, pp. 81–89.
- [20] B. R. Warner and R. Neville-Shepard, "Echoes of a conspiracy: Birthers, truthers, and the cultivation of extremism," *Communication Quarterly*, vol. 62, no. 1, pp. 1–17, 2014.
- [21] W. J. Campbell, *Getting it wrong: ten of the greatest misreported stories in American journalism*. Univ of California Press, 2010.
- [22] G. Aisch, J. Huang, and C. Kang, "Dissecting the# pizzagate conspiracy theories," *The New York Times*, vol. 10, 2016.
- [23] S. Jackson, "Conspiracy theories in the patriot/militia movement," 2017.

- [24] E. Chandrasekharan, U. Pavalanathan, A. Srinivasan, A. Glynn, J. Eisenstein, and E. Gilbert, "You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech," *Proc. ACM Hum.-Comput. Interact.*, vol. 1, no. CSCW, pp. 31:1–31:22, Dec. 2017. [Online]. Available: <http://doi.acm.org/10.1145/3134666>
- [25] M. J. Salganik and D. J. Watts, "Leading the herd astray: An experimental study of self-fulfilling prophecies in an artificial cultural market," *Social psychology quarterly*, vol. 71, no. 4, pp. 338–355, 2008.
- [26] A. Thudt, U. Hinrichs, and S. Carpendale, "The bohemian bookshelf: supporting serendipitous book discoveries through information visualization," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2012, pp. 1461–1470.
- [27] T. C. Schelling, "Dynamic models of segregation," *Journal of mathematical sociology*, vol. 1, no. 2, pp. 143–186, 1971.
- [28] R. Hegselmann, U. Krause *et al.*, "Opinion dynamics and bounded confidence models, analysis, and simulation," *Journal of artificial societies and social simulation*, vol. 5, no. 3, 2002.
- [29] P. Sen and B. K. Chakrabarti, *Sociophysics: an introduction*. Oxford University Press, 2013.
- [30] É. Danchin, L.-A. Giraldeau, T. J. Valone, and R. H. Wagner, "Public information: from nosy neighbors to cultural evolution," *Science*, vol. 305, no. 5683, pp. 487–491, 2004.
- [31] P. Pirolli and S. Card, "Information foraging," *Psychological review*, vol. 106, no. 4, p. 643, 1999.
- [32] J.-L. Deneubourg and S. Goss, "Collective patterns and decision-making," *Ethology Ecology & Evolution*, vol. 1, no. 4, pp. 295–311, 1989.
- [33] M. Belz, L. W. Pyritz, and M. Boos, "Spontaneous flocking in human groups," *Behavioural processes*, vol. 92, pp. 6–14, 2013.
- [34] G. Le Bon, *The crowd: A study of the popular mind*. Fischer, 1897.
- [35] S. Bikhchandani, D. Hirshleifer, and I. Welch, "A theory of fads, fashion, custom, and cultural change as informational cascades," *Journal of political Economy*, vol. 100, no. 5, pp. 992–1026, 1992.
- [36] R. Epstein and R. E. Robertson, "The search engine manipulation effect (seme) and its possible impact on the outcomes of elections," *Proceedings of the National Academy of Sciences*, vol. 112, no. 33, pp. E4512–E4521, 2015.
- [37] M. A. Nowak and K. Sigmund, "Tit for tat in heterogeneous populations," *Nature*, vol. 355, no. 6357, p. 250, 1992.
- [38] M. Nowak and K. Sigmund, "A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game," *Nature*, vol. 364, no. 6432, pp. 56–58, 1993.
- [39] J. Gittins, K. Glazebrook, and R. Weber, *Multi-armed bandit allocation indices*. John Wiley & Sons, 2011.
- [40] M. Bacharach, *Beyond individual choice: teams and frames in game theory*. Princeton University Press, 2006.
- [41] E. Bonabeau, "Agent-based modeling: Methods and techniques for simulating human systems," *Proceedings of the National Academy of Sciences*, vol. 99, no. suppl 3, pp. 7280–7287, 2002.
- [42] C. W. Reynolds, "Flocks, herds and schools: A distributed behavioral model," *ACM SIGGRAPH computer graphics*, vol. 21, no. 4, pp. 25–34, 1987.
- [43] F. Cucker and S. Smale, "Emergent behavior in flocks," *IEEE Transactions on automatic control*, vol. 52, no. 5, pp. 852–862, 2007.
- [44] R. Olfati-Saber, "Flocking for multi-agent dynamic systems: Algorithms and theory," *IEEE Transactions on automatic control*, vol. 51, no. 3, pp. 401–420, 2006.
- [45] G. J. Stephens, L. J. Silbert, and U. Hasson, "Speaker–listener neural coupling underlies successful communication," *Proceedings of the National Academy of Sciences*, vol. 107, no. 32, pp. 14 425–14 430, 2010.
- [46] M. Gallotti, M. Fairhurst, and C. Frith, "Alignment in social interactions," *Consciousness and cognition*, vol. 48, pp. 253–261, 2017.
- [47] J. Kennedy, *Particle Swarm Optimization*. Boston, MA: Springer US, 2010, pp. 760–766. [Online]. Available: https://doi.org/10.1007/978-0-387-30164-8_630
- [48] F. Ye, "Particle swarm optimization-based automatic parameter selection for deep neural networks and its applications in large-scale and high-dimensional data," *PloS one*, vol. 12, no. 12, p. e0188746, 2017.
- [49] K. Giles, "Handbook of russian information warfare," *NATO Defense College, Research Division*, 2016.
- [50] L. G. Stewart, A. Arif, and K. Starbird, "Examining trolls and polarization with a retweet network," 2018.
- [51] M. Stella, M. Cristoforetti, and M. De Domenico, "Influence of augmented humans in online interactions during voting events," *arXiv preprint arXiv:1803.08086*, 2018.
- [52] D. Strömbom, R. P. Mann, A. M. Wilson, S. Hailes, A. J. Morton, D. J. Sumpter, and A. J. King, "Solving the shepherding problem: heuristics for herding autonomous, interacting agents," *Journal of the royal society interface*, vol. 11, no. 100, p. 20140719, 2014.
- [53] D. H. Patent, *The buffalo and the Indians: A shared destiny*. Houghton Mifflin Harcourt, 2006.
- [54] A. Kydd, "Sheep in sheep's clothing: Why security seekers do not fight each other," *Security Studies*, vol. 7, no. 1, pp. 114–155, 1997.
- [55] V. Gerasimov, "The value of science is in the foresight: New challenges demand rethinking the forms and methods of carrying out combat operations," *Military Review*, vol. 96, no. 1, p. 23, 2016.
- [56] M. Boguñá, R. Pastor-Satorras, A. Díaz-Guilera, and A. Arenas, "Models of social networks based on social distance attachment," *Physical review E*, vol. 70, no. 5, p. 056122, 2004.
- [57] V. Schwämmle, M. González, A. Moreira, J. S. Andrade Jr, and H. Herrmann, "Different topologies for a herding model of opinion," *Physical Review E*, vol. 75, no. 6, p. 066108, 2007.
- [58] K. E. Iverson, "A programming language," in *Proceedings of the May 1-3, 1962, Spring Joint Computer Conference*, ser. AIEE-IRE '62 (Spring). New York, NY, USA: ACM, 1962, pp. 345–351. [Online]. Available: <http://doi.acm.org/10.1145/1460833.1460872>
- [59] R. Lande, "Models of speciation by sexual selection on polygenic traits," *Proceedings of the National Academy of Sciences*, vol. 78, no. 6, pp. 3721–3725, 1981.
- [60] S. Salvador and P. Chan, "Toward accurate dynamic time warping in linear time and space," *Intelligent Data Analysis*, vol. 11, no. 5, pp. 561–580, 2007.
- [61] R. Bellman, *Dynamic programming*. Courier Corporation, 1957.
- [62] J. D. Cohen, S. M. McClure, and J. Y. Angela, "Should i stay or should i go? how the human brain manages the trade-off between exploitation and exploration," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 362, no. 1481, pp. 933–942, 2007.
- [63] M. Granovetter, "Threshold models of collective behavior," *American journal of sociology*, vol. 83, no. 6, pp. 1420–1443, 1978.
- [64] C. R. Sunstein, "The law of group polarization," *Journal of political philosophy*, vol. 10, no. 2, pp. 175–195, 2002.
- [65] W. Quattrociocchi, "Inside the echo chamber," *Scientific American*, vol. 316, no. 4, pp. 60–63, 2017.
- [66] H. Tajfel and J. C. Turner, "The social identity theory of intergroup behavior," 2004.
- [67] P. André, J. Teevan, S. T. Dumais *et al.*, "Discovery is never by chance: designing for (un) serendipity," in *Proceedings of the seventh ACM conference on Creativity and cognition*. ACM, 2009, pp. 305–314.
- [68] C. Lamay, "Public service advertising, broadcasters, and the public interest," *Shouting to be heard: Public service advertising in a new media age*, 2002.
- [69] S. A. West, A. S. Griffin, and A. Gardner, "Evolutionary explanations for cooperation," *Current Biology*, vol. 17, no. 16, pp. R661–R672, 2007.
- [70] S. R. Dall, L.-A. Giraldeau, O. Olsson, J. M. McNamara, and D. W. Stephens, "Information and its use by animals in evolutionary ecology," *Trends in ecology & evolution*, vol. 20, no. 4, pp. 187–193, 2005.
- [71] A. Kirilenko, A. S. Kyle, M. Samadi, and T. Tuzun, "The flash crash: High-frequency trading in an electronic market," *The Journal of Finance*, 2017.
- [72] J. J. Thomson, "Killing, letting die, and the trolley problem," *The Monist*, vol. 59, no. 2, pp. 204–217, 1976.
- [73] M. R. Waser, "Discovering the foundations of a universal system of ethics as a road to safe artificial intelligence," in *AAAI Fall Symposium: Biologically Inspired Cognitive Architectures*, 2008, pp. 195–200.
- [74] B. Deng, "The robot's dilemma," *Nature*, vol. 523, no. 7558, p. 24, 2015.
- [75] L. Nelson, "Firefighters attempt to contain bel-air blaze ahead of the strong winds expected thursday night," Dec 2017. [Online]. Available: <https://tinyurl.com/pishkinFire>