Destabilising Conventions using Temporary Interventions

James Marchant, Nathan Griffiths, Matthew Leeke, and Henry Franks

Department of Computer Science, University of Warwick, Coventry, UK
{james, nathan, matt}@dcs.warwick.ac.uk, hpwfranks@googlemail.com

Abstract. Conventions are an important concept in multi-agent systems as they allow increased coordination amongst agents and hence a more efficient system. Encouraging and directing convention emergence has been the focus of much research, particularly through the use of fixed strategy agents. In this paper we apply temporary interventions using fixed strategy agents to destabilise an established convention by (i) replacing it with another convention of our choosing, and (ii) allowing it to destabilise in such a way that no other convention explicitly replaces it. We show that these interventions are effective and investigate the minimum level of intervention needed.

Keywords: Convention emergence · Norms · Coordination · Intervention

1 Introduction

In multi-agent systems (MAS) coordinated actions help to reduce the costs associated with incompatible choices and increase the efficiency of a system. However, in many domains such behaviour cannot be enforced, as there is no centralised control and a lack of *a priori* knowledge of which actions clash. In practice, many systems rely on the evolution of *conventions* as standards of behaviour adopted by agents with no, or little, involvement from system designers. Understanding how these conventions emerge, how they can be influenced, and how aspects such as topology affect them is an active research area [5, 7, 10, 15, 18].

Conventions have been shown to support high levels of coordination without the need to dictate action choices in a top-down manner. Facilitating the emergence of high-quality conventions in a short period of time, without requiring prior computation, is an area of ongoing research. Much work has focussed on the emergence of conventions given only agent rationality and the ability to learn from previous choices. Small numbers of fixed strategy agents (agents who choose the same action regardless of others' choices) have been shown to influence the conventions that emerge and to increase the speed of adoption [7, 8, 15].

The ability to remove, as well as establish, conventions allows correction or replacement of adopted actions. In domains where the desirability of actions can change over time, being able to cause such a change is beneficial to the system as

a whole. Additionally, understanding how to cause this shift gives insights into what makes a convention robust to outside influence.

In this paper, we examine what is needed to *destabilise* an established convention. We propose temporarily inserting agents, known as *Intervention Agents* (IAs), with strategies that differ from the established convention to influence a population into discarding that convention. The insertion of IAs is equivalent to incentivising individuals to take particular actions, for example through reward or payment. We show that a small proportion of IAs placed at targeted locations in the population for a sufficient duration can destabilise an established convention, replacing it with another of our choosing. We also show that conventions can be destabilised in such a way that we are not required to select a replacement, and instead we can allow a new convention to emerge.

The remainder of this paper is structured as follows. In Section 2 we introduce the related work on convention emergence and the role of fixed strategy agents. Sections 3 and 3.1 present our model of convention emergence and metrics for characterising conventions. In Section 3.2 we present our model of IAs for convention destabilisation. We describe our experimental settings in Section 4, and present our results in Section 5. Finally, in Section 6 we present our conclusions.

2 Related Work

Conventions can be viewed as socially-accepted rules, in the form of expected behaviour, amongst agents. There is no explicit punishment for acting against the convention, but doing so increases the likelihood of coordination problems and costs. Thus a convention can be thought of as "an equilibrium everyone expects in interactions that have more than one equilibrium" [20]. Conventions can emerge from local agent interactions [5, 10, 17, 19] and support coordination by placing *social constraints* on the actions that are available to the agents [16]. As such, conventions differ from *norms* (although the terms are often used interchangeably in the literature [12, 15]) as the latter typically involve punishments for failure to adhere to the expected behaviour [2, 3, 9, 14]. Norms generally require additional abilities or overheads to facilitate this punishment. We do not assume that agents are able to punish others (or even to observe their defection), and instead focus on conventions as a lightweight method of supporting coordination.

In this work we examine convention emergence where the only assumptions on agent behaviour are rationality and a (limited) *memory* of past interactions. This setting has been widely studied [5, 8, 15, 19] and is able to support effective convention emergence. Walker and Wooldridge [19] were amongst the first to produce a formal model of convention emergence with few assumptions about the underlying agent architecture. They present a model in which a global convention emerges where agents choose their action based solely on observations of others. Sen and Airiau [15] explore social learning as a method for convention emergence, where agents learn the best action choice based on the payoff of their interactions. They show that convention emergence is possible with min-

imal additions to agents' abilities (for example, no memory of interactions is required) and without assuming public interactions. However, the work is limited by several simplifications: there is no connecting topology restricting agent interactions and the convention space contains only two possible conventions. In general, larger convention spaces and connecting topologies are commonplace.

The underlying topology has been shown to have a significant effect on convention emergence [4, 5, 10, 18]. Much of the work investigating topology has been restricted to a small convention space (typically with just two actions). More recent work has explored the effect of increasing the number of available actions and has shown that doing so typically increases the time taken for convergence [7, 8, 13].

2.1 Fixed Strategy Agents

Sen and Airiau [15] demonstrated that a small number of fixed strategy agents, that always choose the same action regardless of others' actions, were sufficient to cause a population to adopt this action as a convention. This indicates that, at least in some circumstances, small numbers of agents are able to influence much larger populations. Franks et al. [6,7] examined the effectiveness of fixed strategy agents when agent interactions are restricted by a social network topology in a large convention space. They showed that the topology affects the number of fixed strategy agents required to influence convergence speed, and that *where* such agents are placed is crucial to the extent of their influence. Placement by metrics such as degree or eigenvector centrality has substantial benefits over random placement on speeding up convergence.

2.2 Destabilisation of Conventions

There has been relatively little work that explores destabilising established conventions. Previous work on fixed strategy agents focuses on promoting convention emergence, by introducing such agents at the beginning of population modelling. Our hypothesis is that fixed strategy agents can also be used to destabilise existing conventions. Villatoro et al. [17,18] explored a similar concept of destabilisation as part of convention emergence. They consider meta-stable subconventions, which are secondary conventions amongst subsets of the population that persist due to their stability. Meta-stable subconventions impede the emergence of more general conventions and can prevent full adoption. Villatoro et al. describe methods for preventing and removing meta-stable subconventions by identifying and targeting particular topological structures. Although related to our work, their approach focuses on subgroups of the population whereas we focus on the whole population. Moreover, Villatoro et al. have the aim of destabilising meta-stable subconventions to enable full emergence of a single convention, while our aim is more broadly to destabilise existing conventions.

3 Convention Emergence Model

Conventions emerge as a result of agents in a population selecting the same action and learning the best strategy (action choice) over time. We assume that a population consists of a set of agents, $Ag = \{1, ..., N\}$, who select from a number of actions, $\Sigma = \{\sigma_1, \sigma_2, ..., \sigma_n\}$. Each timestep each agent selects an interaction partner from its neighbours, and both partners choose an action from Σ. The individual payoff for each agent is determined by the combination of action choices. In this paper we adopt the n-action coordination game, such that interaction partners receive a positive payoff if they select the same action and a negative payoff if their actions differ. The 2-action coordination game is often used in exploring convention emergence, but we expand to the n-action coordination game to avoid restricting the number of possible conventions as discussed above.

Each agent chooses the action that it believes will result in the highest utility based on its previous interactions. We also assume an element of exploration, such that with probability $p_{explore}$ agents will choose a random action from those available. In this regard our model adopts the approach of Villatoro et al. [18] by using a simplified Q-Learning algorithm for both partners in an interaction to update their strategies.

We assume that agents are situated on a topology that restricts their interactions such that agents can only interact with their neighbours. Further, we consider small-world and scale-free networks which exhibit properties that reflect those observed in real-world environments such as power law degree distributions and clustering. We also consider random networks as a baseline.

3.1 Convention Metrics

In order to characterise convention establishment we need a measure of when a convention exists and when agents should be considered as members of that convention. Much work in the field uses Kittock's criteria in which a convention is said to have emerged when 90% of the non-fixed strategy agents, when not exploring, select the same action [10]. However this offers no insights into emerging conventions until after they have become established, or of their decline if they are subsequently destabilised. Additionally, this measure relies on observation of agent internals to know when they are exploring and their preferred action. Thus, we propose a finer grained set of metrics for characterising convention emergence, from which we will define our strategies for destabilisation.

We introduce a number of new metrics (modified from [19]). We begin by formalising what it means to say an agent chose an action:

$$chose_x(\sigma, t) \iff \exists i : i \in par_x(t) \wedge self_x(i, t) = \sigma \tag{1}$$

where $self_x(i, t)$ is the action chosen by agent x in interaction i in timestep t, and $par_x(t)$ is the set of interactions that x participated in during timestep t.

We can then define the set of agents that have chosen a given action $\sigma \in \Sigma$ during timestep t as:

$$chosen(\sigma, t) = \{x | x \in Ag \wedge chose_x(\sigma, t)\} \tag{2}$$

We also require a way of defining whether we consider an agent to be a member of a convention or not, and of establishing the *existence* of a convention. Due to exploration, full adherence to a single strategy is unlikely to occur. It is useful to quantify an agent's *adherence* to a strategy of choosing ϕ as the probability of that agent choosing ϕ in any potential interaction at time t:

$$adh(x, \phi, t) = P(self_x(i, t) = \phi \mid i \in par_x(t)) \tag{3}$$

Note that since in general action selection is likely to be relatively complex, we may not be able to establish adherence exactly. We can determine an estimate based on the agent's interaction history, by considering the proportion of the last λ interactions in which the agent selected ϕ.

We subsequently define the set of conventions Φ_t that exist in a population at time t as follows:

$$\phi \in \Phi_t \iff \exists x : x \in chosen(\phi, t) \wedge adh(x, \phi, t) > \gamma \tag{4}$$

That is, a given action σ is considered to be a convention at time t if there is at least one agent choosing that action with a probability greater than some threshold γ. This characterisation allows us to capture the notion of a personal convention analogous to that of a personal norm. We use ϕ to denote an action that is also a convention and σ to denote an action that may or may not be a convention. This distinction allows us to separate actions selected by chance, exploration or some other process and those selected with sufficient frequency to be considered conventions.

We define the average adherence to a strategy of choosing σ to be the mean adherence across the agents that chose σ in a timestep:

$$averageAdh(\sigma, t) = \frac{\sum_{x \in chosen(\sigma, t)} adh(x, \sigma, t)}{|chosen(\sigma, t)|} \tag{5}$$

We assume that the temporal variance of adh is low, such that an agent who satisfies $adh(x, \phi, t) > \gamma$ at time t is likely to satisfy it at $t + 1$ (Walker and Wooldridge [19] discussed that since strategy change typically incurs a cost we can expect the number of strategy changes to be minimised).

We define a convention as *established* if the average adherence is greater than the *convention establishment threshold* β, a model-wide parameter:

$$estbl(\phi, t) \iff \phi \in \Phi_t \wedge averageAdh(\phi, t) > \beta \tag{6}$$

Finally, we can define the extent to which agents are part of a convention. We denote agents as *members* of a convention if they currently adhere to it with probability greater than or equal to β:

$$member(x, \phi, t) \iff estbl(\phi, t) \wedge adh(x, \phi, t) \geq \beta \tag{7}$$

Thus, the membership set for a given convention at time t is given by:

$$membership(\phi, t) = \{x | x \in Ag, \phi \in \Phi_t, member(x, \phi, t)\} \quad (8)$$

By measuring the size of convention membership sets over time we can monitor how conventions become established and grow without internal observation of agents' decision making. Furthermore, we can distinguish between agents who used a convention due to exploration and those who are truly members.

3.2 Intervention Agents

As discussed in Section 2, fixed strategy agents can influence convention emergence when introduced at the beginning of a simulation. We call these fixed strategy agents *Intervention Agents* (IAs) and, unlike in previous work, they are introduced to destabilise established conventions. Building on the work of Franks et al. [6,7] we propose simultaneously introducing IAs to replace nodes from the primary convention (that with the highest membership) to manipulate convention emergence. The duration of IA placement is varied to investigate the extent of intervention required to elicit a lasting change on the primary convention.

There are two types of destabilisation we can attempt using IAs: *aggressive* and *non-aggressive*. In aggressive destabilisation the aim is to *demote* the primary convention and *promote* a specified alternative convention in its place. In our experimentation we select the second most adopted convention as the alternative for promotion. Thus, we use IAs to encourage members of the primary convention to adopt the secondary convention. Non-aggressive destabilisation aims to demote the primary convention without having to select an alternative convention in its place, instead allowing a new convention to emerge. To accomplish this we propose that IAs adopt a uniform distribution of actions selected from those not already established as conventions. Our hypothesis is that this will destabilise the primary convention and allow an alternative to emerge.

4 Experimental Setup

We performed experiments with populations of 1000 agents, that use Q-learning (with a learning rate and an exploration rate of 0.25) to evolve their strategies. Unless otherwise stated we use the 10-action coordination game. We explored other sizes of convention space and obtained similar results to those presented here. All results are averaged over 30 runs, unless otherwise stated.

A window size of $\lambda = 30$ is used for adherence approximation, giving sufficient granularity to estimate membership whilst minimising memory overhead. The required action selection probability for an action to be considered a convention, γ, is 0.5. This enables more strategies to be considered as conventions (whether or not they are established) to give more information on the effects of intervention.

The convention emergence threshold, β, is set to 0.9 (in line with other work as discussed above). However, due to our method of measuring convention emergence we do not assume knowledge of whether an agent is exploring. As such,

the 90% threshold must be adjusted to take into account the random exploration of agents (noting that when exploring the agent can potentially still choose the "best" action $1/N$ times). This gives: $\beta = 0.9 \times (1 - (p_{explore}(N-1))/N))$, where N is the number of actions, $p_{explore}$ is the exploration rate, and $(N-1)/N$ represents the ratio of randomly chosen actions that are not the "best".

We used the Java Universal Network/Graph Library (version 2.0.1)[1] to generate interaction topologies. Scale-free topologies were generated using the Barabási-Albert algorithm with parameters $m_0 = 4$, $m = 3$, where m_0 is the initial number of vertices and $m \leq m_0$ is the number of edges added from a new node to existing nodes each evolution [1]. Small-world topologies were generated using the Kleinberg model with a lattice size of 10×100, clustering exponent $\alpha = 5$ and one "long-distance" connection per node [11].

We ran simulations for 5000 timesteps before introducing IAs, since this was found to be sufficient time for convention emergence and stabilisation in all topologies. At timestep 5000 a set of IAs were introduced, replacing nodes within the primary convention selected either randomly or by highest degree. The IAs remain for either a fixed number of timesteps or until the end of the simulation, to investigate the duration required for destabilisation and whether the primary convention can recover when the IAs are removed. Upon removal of the fixed strategy nature agents again use Q-learning to choose actions (with learning continuing during the fixed strategy period). Unless otherwise stated, the simulations were performed for 10000 iterations in total, to enable replacement conventions to emerge after destabilisation.

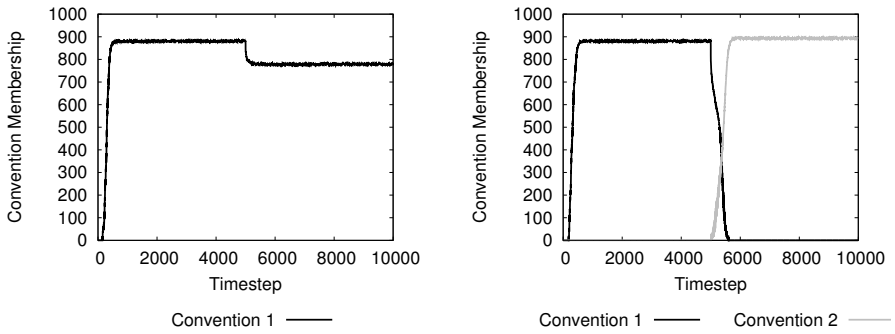
If there are insufficient members of the primary convention for the target number of IAs to be introduced then additional IAs are placed throughout the rest of the population according to the current placement strategy. Note that this implies the primary convention is immediately destabilised (as all of its members are now IAs) but such settings are included for completeness.

5 Results and Discussion

5.1 Number of fixed strategy agents

We begin by examining the effect of introducing a varying number of IAs into the population indefinitely. To establish a baseline for the minimum number of IAs required to enact a change, we introduce a set of IAs at time 5000 that remain until the end of the simulation. For these results we use *aggressive destabilisation*, such that IAs use the action of the secondary convention (determined by ranking conventions by membership size and then average adherence). IAs replace the highest degree agents that were members of the primary convention. We also performed experiments using random placement, which confirmed the results of Franks et al. that random placement is inferior to placement by degree [6, 7]. Thus, in the remainder of this paper we focus on placement by degree.

[1] http://jung.sourceforge.net/



(a) 50 Intervention Agents (b) 100 Intervention Agents

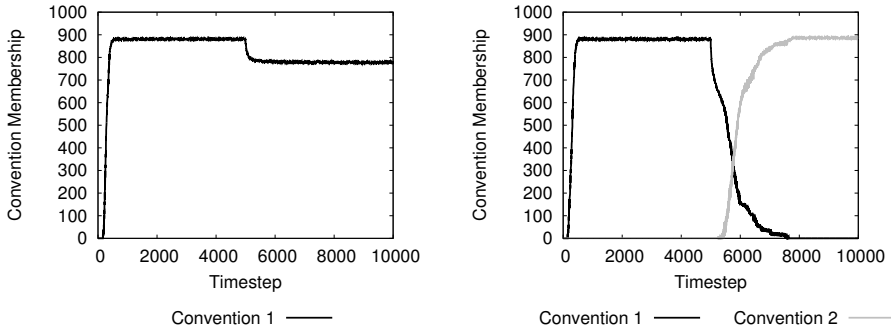
Fig. 1. The effect of Intervention Agents on random graphs

Our simulations were performed on scale-free and small-world networks as described in Section 4, and on random graphs generated using the Erdös-Rényi generator to provide a baseline. In order to provide similar edge numbers to the scale-free and small-world graphs we used a connection probability of 0.006.

For all figures in this section (excluding Figure 8), all conventions which have significant non-zero membership during the simulation are plotted. Those with zero or near-zero membership have been excluded from the graphs for clarity.

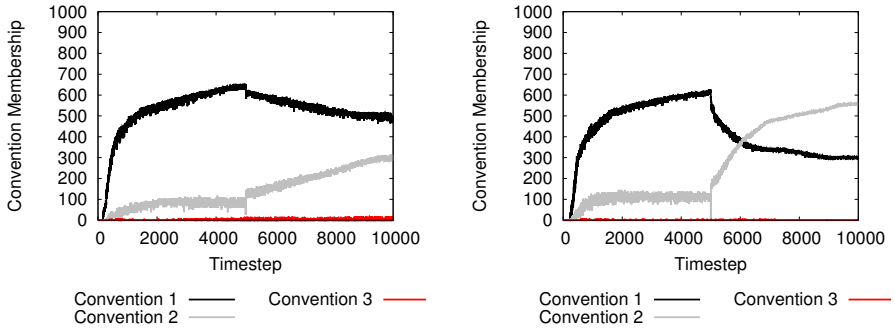
Figure 1 shows the effect of different numbers of IAs on the random graph topology. With the introduction of 50 IAs the membership of the primary convention (displayed in black) drops (more so than the 50 agents who became IAs would account for) but stabilises again rather than destabilising completely. This is likely due to the IAs being able to influence a local area around themselves but agents further away being too adherent to the primary convention to be affected. We would therefore expect this "dip" to increase in depth as the number of IAs increases and, indeed, this is what was observed. This behaviour continued until around 80-100 IAs after which the primary convention becomes destabilised enough for the secondary convention (displayed in grey) to overtake it. This behaviour is shown in Figure 1b. Of particular interest is that the speed with which the changeover happens indicates that, once the critical number of IAs are included, they are only needed for a short period of time.

Figure 2 shows results for scale-free graphs. We see similar behaviour to random graphs, with the decrease in membership of the primary convention increasing proportionally to the number of IAs until a critical number of IAs where the destabilisation is enough to allow the promotion of the secondary convention. Scale-free networks require significantly fewer IAs than random graphs, needing only 40, to achieve destabilisation. This is accounted for by the presence of "hubs" which are able to influence large groups of agents, at least some of which will be chosen as locations for IAs due to their high degree. In both cases however the primary convention is fully destabilised whilst the secondary is promoted to the same membership size as the primary originally had.



(a) 20 Intervention Agents (b) 40 Intervention Agents

Fig. 2. The effect of Intervention Agents on scale-free graphs

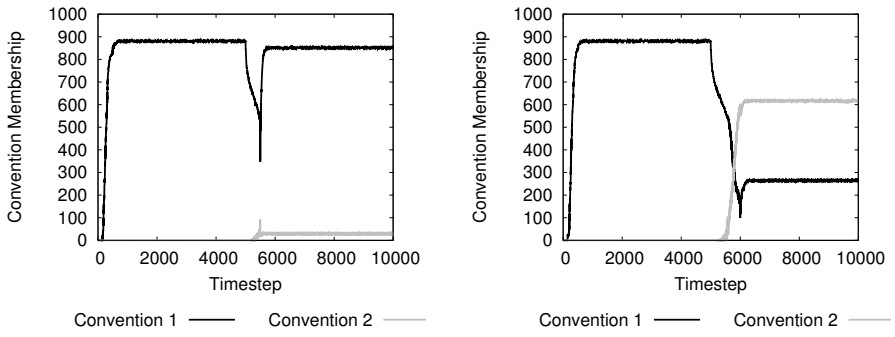


(a) 20 Intervention Agents (b) 40 Intervention Agents

Fig. 3. The effect of Intervention Agents on small-world graphs

Results for small-world networks are shown in Figure 3. Whilst the overall behaviour is similar, in that there is a critical number of IAs after which destabilisation will occur, the behaviour pre-transition is less well-defined. In particular, the characteristic "dip" that occurs in scale-free and random topologies is not present to the same extent, and the drop in membership of the primary convention is slower. Additional simulations over longer durations show that the convention does eventually stabilise but takes a large number of iterations (approximately 20000). This likely follows from the clustered nature of small-world graphs, and we hypothesise that the clusters are slow to adapt to the changes in convention. This hypothesis is supported by the number of agents in the primary convention before intervention being substantially lower than in scale-free and random graphs, implying that the clustering slows convention emergence. Previous work by Franks et al. observed similar disparities in convention adoption between scale-free and small-world graphs [7].

Full destabilisation, as seen in Figures 1 and 2, was found to occur in small-world topologies with 70 or more fixed agents, with 40 the minimum required



(a) 500 Timesteps (b) 1000 Timesteps

Fig. 4. The effect on scale-free graphs of 40 IAs when introduced for finite time

to replace the primary convention. Additionally, a third convention is present at the bottom of the graphs. The presence of this is unique to small-world graphs and is included to show this difference between topologies.

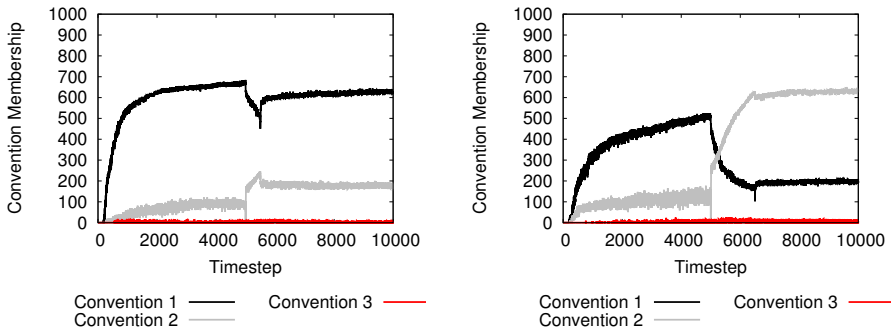
5.2 Length of Intervention

We have shown that destabilisation is possible and have identified the smallest number of IAs needed. In this section we determine the duration needed for a permanent change, by adding IAs temporarily for a fixed duration.

Figure 4 shows the effect of varying the duration for which 40 IAs are present on scale-free graphs. In the previous section we demonstrated that 40 IAs is sufficient for destabilisation. In Figure 4a IAs are introduced for 500 timesteps. Whilst the characteristic decrease in numbers we saw previously starts to occur, when IAs are removed the primary convention rapidly recovers. However, we begin to see the effect of IAs since after the intervention a stable secondary convention emerges. The size of this convention is comparable to the difference in the primary convention size before and after intervention, implying that the second convention represents agents who have permanently changed convention.

Increasing the intervention length to 1000 timesteps, as shown in Figure 4b, is sufficient for destabilisation, and for the secondary convention to overtake the primary. However, it is not as well established as with permanent interventions, indicating that keeping the IAs for longer would further destabilise the primary convention. This was verified by testing over longer time periods. It is also worth noting that the primary convention manages to recover slightly before stabilising, but that this does not shrink the secondary convention. Therefore, the primary convention is regaining nodes that were no longer strong adherents to the primary but had not yet become strong adherents to the secondary convention.

Figure 5 shows the effect of temporary interventions for small-world graphs. Figure 5a shows similar behaviour to its scale-free counterpart as the intervention duration is insufficient for destabilisation. The change in membership is larger



(a) 500 Timesteps (b) 1500 Timesteps

Fig. 5. The effect on small-world graphs of 40 IAs when introduced for finite time

than in scale-free networks, in terms of absolute and relative size, which supports our hypothesis regarding the clustered nature of small-world graphs implying that influence internal to a cluster is easier than changing it externally. As such, when clusters change from the primary to the secondary convention they are unlikely to change back when IAs are removed.

Figure 5b shows that the length of intervention needed for permanent change in small-world networks is longer than for scale-free networks, taking 1500 iterations rather than 1000. This is due to the clustered nature of small-world graphs, and supports the findings of Griffiths and Anand [8] showing that small-world networks converge slower than scale-free and, in this case, take longer to change.

5.3 Non-Aggressive Destabilisation

Previous simulations have focused on aggressive destabilisation, where the primary convention is demoted whilst promoting the secondary. We now consider *non-aggressive destabilisation* where we attempt to destabilise the primary convention without explicitly promoting another convention in its place.

Figure 6 shows sample runs from inserting 100 IAs that replace the high degree nodes of the primary convention in a scale-free topology. In Figures 6a and 6b the IAs are inserted indefinitely, while in 6c and 6d they are removed after 2000 iterations. Unlike aggressive destabilisation the IA strategies are selected uniformly at random from the bottom 7 ranked strategies at time 5000. Each plot shows a different run, since average results are not appropriate as the final emergent convention differs at random. The runs show the same behaviour, with the primary convention being destabilised around timestep 6000. This is slower than the destabilisation achieved with aggressive IAs, but is expected due to the lack of a coordinated effort to replace the influenced agents' strategies. In each run a new convention emerges around timestep 8500 and since this convention emerges naturally it may differ each time. By destabilising the primary convention, but not explicitly favouring another, a new convention naturally

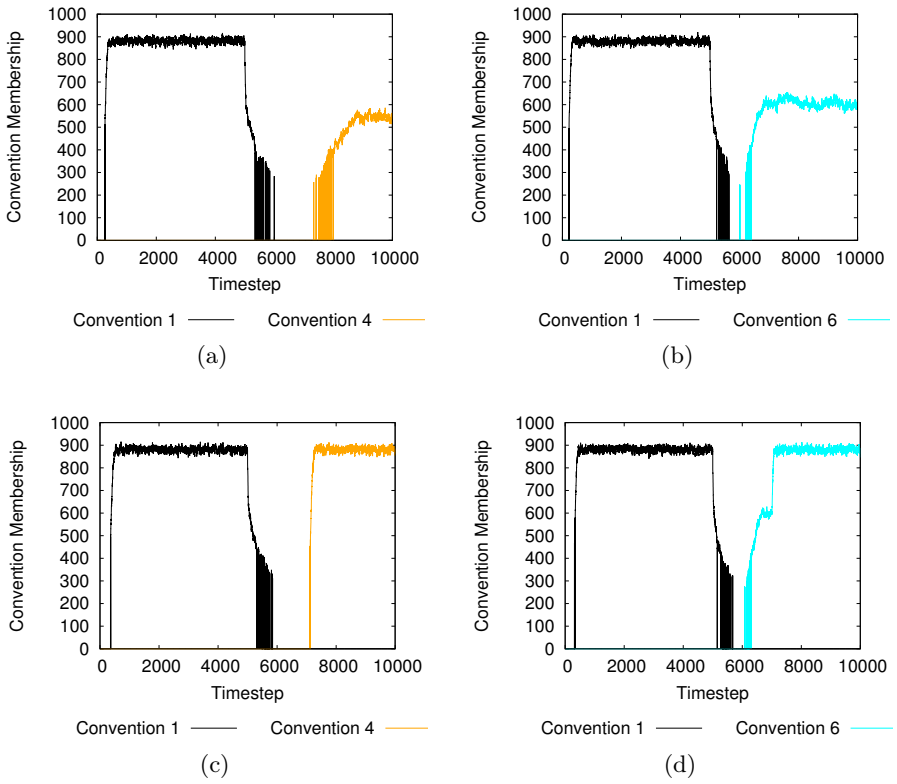


Fig. 6. Non-aggressive destabilisation in scale-free graphs. In (a) and (b) the IAs remain indefinitely. In (c) and (d) they remain for 2000 timesteps.

emerges, but we cannot predict what it will be. This contrasts with aggressive destabilisation where the secondary (targeted) convention always emerges.

The length of the intervention affects the final membership size attained by the new convention. Where IAs remain indefinitely (the upper two plots) the stable membership size is several hundred less than with a temporary intervention (the lower two plots). This can be explained by the presence of the IAs, which randomly select strategies from the lowest 7 conventions at the time of initial intervention, continuing to hinder the new convention from spreading in much the same way as they destabilised the original primary convention. We would expect that when IAs are removed the new convention will spread to the area that they were occupying, which is seen in Figure 6 (lower two plots) where the new convention undergoes rapid size increase as soon as the IAs are removed.

Figure 7 shows individual non-aggressive runs on small-world topologies. The length of time the IAs are present is the same as in Figure 6 but the number of IAs is increased to 200, as 100 agents is insufficient for destabilisation in this setting. This relates to the hypothesis that the clusters in small-world topologies

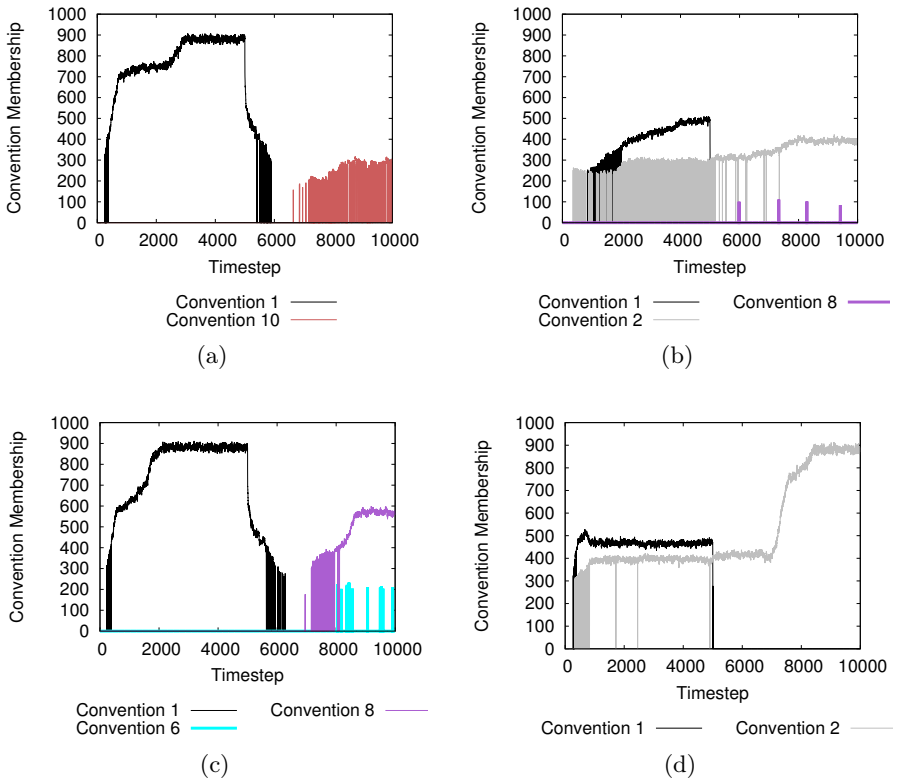
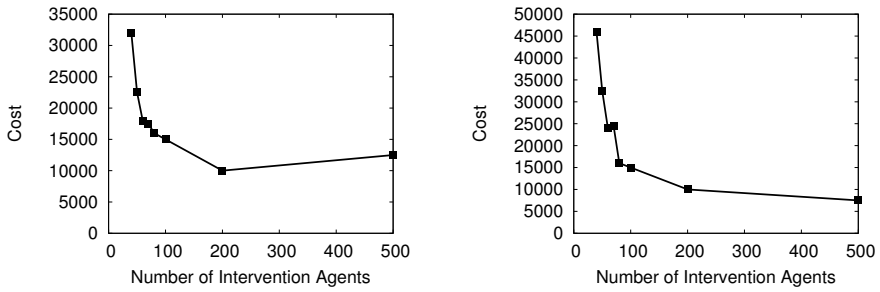


Fig. 7. Non-aggressive destabilisation in small-world graphs. In (a) and (b) the IAs remain indefinitely. In (c) and (d) they remain for 2000 timesteps.

make change slower and more difficult to achieve than in scale-free topologies. The findings are similar to those in scale-free graphs: once destabilisation of the primary convention has occurred another emerges after a small period of time and without predictable strategy. Whilst the overall convention membership size is smaller than those in scale-free topologies (as was the case in other simulations) the relationship between the IAs remaining and the lower membership sizes of the replacement convention still holds. Again, removing IAs after destabilisation is conducive to a stronger new convention emerging.

5.4 Cost of Intervention vs. Effect

Finally, we consider the relationship between the number of IAs and the duration of intervention needed for destabilisation. We define one unit of cost to be one IA being included for one iteration. Thus, 200 IAs present for one iteration have a cost of 200, while 5 IAs present for 20 iterations have a cost of 100. Since real-world interventions, such as incentives or payments, are likely to have a tangible cost it is useful to measure the expense of a strategy for using IAs.



(a) Scale-free (b) Small-world

Fig. 8. Number of IAs vs. the minimum cost to cause destabilisation

In this set of experiments we varied the number of IAs from 40 (the minimum number required for destabilisation) to 500, while simultaneously increasing the duration of intervention from 0 in steps of 50 until destabilisation occurred. For numbers of IAs above 200, where more granularity in the length of time was needed, the duration was increased in steps of 5. These values were then used to calculate the minimum cost associated with causing the destabilisation.

For both small-world and scale-free topologies, increasing the number of agents decreased the cost needed for destabilisation, as shown in Figure 8. This is because, even though the number of IAs increases, the required duration decreases by a higher proportion, resulting in a lower cost. The effect of increasing the number of agents is one of diminishing returns: increasing the number of agents produces smaller reductions in cost each time. In addition, whilst influencing small numbers of agents in a population is likely to be possible, being able to influence half of all agents is, due to the lack of centralised control, unlikely in most domains, and so 500 IAs are included only for completeness.

Whilst the relationship between cost and number of IAs is similar in scale-free and small-world networks it is worth noting that the costs associated with intervening in a small-world topology are substantially higher than those for scale-free topologies. This is due to the need to include IAs for longer periods in small-world topologies, which relates to the decreased speed with which small-world graphs allow conventions to emerge. However, these results show that in general as many IAs as possible should be introduced if they require an ongoing cost. If, instead, placing them only requires a one-off cost, then using the minimum number for destabilisation is preferable as additional agents will increase the cost with little additional benefit (as can be seen in previous sections).

6 Conclusions

We have shown that it is possible to destabilise established conventions by introducing a small proportion of IAs. When using aggressive IAs, whose fixed strategy is that of the second most popular convention, to replace the highest

degree nodes in the primary convention we found that 40 agents (4% of the population) is sufficient to destabilise the primary convention and for the secondary convention to be promoted. This occurs in small-world, scale-free and random topologies, with the latter requiring 100 agents for destabilisation to occur.

We also investigated the minimum duration that IAs must remain in order to cause a permanent destabilisation and prevent the primary convention from re-establishing itself. We found that there was a minimum number of agents and a minimum duration needed to cause this effect, and that the minimum duration for small-world graphs is longer than that required for scale-free graphs. Interventions less than this minimum duration cause a temporary decrease in membership of the targeted convention, which disappears when IAs are removed.

A different method of destabilisation was investigated in the form of non-aggressive destabilisation, which attempts to demote the primary convention without explicitly promoting another. We found that the number of IAs required was higher than in aggressive destabilisation, and that small-world topologies required more IAs than scale-free topologies. We showed that the primary convention would be destabilised and that, whilst a new convention would emerge, its strategy was unpredictable.

Finally, we proposed a method of calculating a "cost" for an intervention and showed that increasing the number of agents was beneficial, assuming that the intervention had an ongoing cost per iteration. We also found that performing interventions was more expensive in small-world than in scale-free topologies.

Overall, we have found that the ability to intervene in a system and remove previously established conventions is possible. The ability to do this means that undesirable conventions can be removed even if they are heavily adhered to, allowing the system to replace such conventions either with direction to a particular convention (the aggressive approach) or through natural emergence.

References

1. Albert, R., Barabási, A.L.: Statistical mechanics of complex networks. Reviews of Modern Physics 74(1), 47–95 (2002)
2. Axelrod, R.: An evolutionary approach to norms. American Political Science Review 80, 1095–1111 (1986)
3. Bicchieri, C., Jeffrey, R.C., Skyrms, B.: The Dynamics of Norms. Cambridge University Press (1997)
4. Delgado, J.: Emergence of social conventions in complex networks. Artificial Intelligence 141(1–2), 171–185 (2002)
5. Delgado, J., Pujol, J.M., Sangüesa, R.: Emergence of coordination in scale-free networks. Web Intelligence and Agent Systems 1(2), 131–138 (2003)
6. Franks, H., Griffiths, N., Anand, S.: Learning agent influence in MAS with complex social networks. Autonomous Agents and Multi-Agent Systems 28(5), 836–866 (2014)
7. Franks, H., Griffiths, N., Jhumka, A.: Manipulating convention emergence using influencer agents. Autonomous Agents and Multi-Agent Systems 26(3), 315–353 (2013)

8. Griffiths, N., Anand, S.S.: The impact of social placement of non-learning agents on convention emergence. In: 11th International Conference on Autonomous Agents and Multiagent Systems. vol. 3, pp. 1367–1368. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC (2012)
9. Kandori, M.: Social norms and community enforcement. The Review of Economic Studies 59(1), 63–80 (1992)
10. Kittock, J.: Emergent conventions and the structure of multi-agent systems. In: Lectures in Complex Systems: the Proceedings of the 1993 Complex Systems Summer School. pp. 507–521. Addison-Wesley, Reading, MA (1995)
11. Kleinberg, J.: Navigation in a small world. Nature 406(6798), 845–845 (2000)
12. Mukherjee, P., Sen, S., Airiau, S.: Norm emergence with biased agents. International Journal of Agent Technologies and Systems 1(2), 71–84 (2009)
13. Salazar, N., Rodriguez-Aguilar, J.A., Arcos, J.L.: Robust coordination in large convention spaces. AI Communications 23(4), 357–372 (2010)
14. Savarimuthu, T.B.R., Arulanandam, R., Purvis, M.: Aspects of active norm learning and the effect of lying on norm emergence in agent societies. In: Kinny, D., Hsu, J.Y.j., Governatori, G., Ghose, A.K. (eds.) Agents in Principle, Agents in Practice, LNCS, vol. 7047, pp. 36–50. Springer, Heidelberg (2011)
15. Sen, S., Airiau, S.: Emergence of norms through social learning. In: 20^{th} International Joint Conference on AI. pp. 1507–1512. Morgan Kaufmann, San Francisco (2007)
16. Shoham, Y., Tennenholtz, M.: On the emergence of social conventions: modeling, analysis, and simulations. Artificial Intelligence 94(1–2), 139–166 (1997)
17. Villatoro, D., Sabater-Mir, J., Sen, S.: Social instruments for robust convention emergence. In: 22^{nd} International Joint Conference on AI. pp. 420–425. AAAI Press (2011)
18. Villatoro, D., Sen, S., Sabater-Mir, J.: Topology and memory effect on convention emergence. In: 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology. vol. 2, pp. 233–240. IEEE Computer Society, Washington, DC (2009)
19. Walker, A., Wooldridge, M.: Understanding the emergence of conventions in multi-agent systems. In: International Conference on Multi-Agent Systems. pp. 384–389. MIT Press (1995)
20. Young, H.P.: The economics of convention. The Journal of Economic Perspectives 10(2), 105–122 (1996)