Destabilising Conventions: Characterising the Cost

James Marchant, Nathan Griffiths and Matthew Leeke
Department of Computer Science
University of Warwick
Coventry, UK, CV4 7AL
Email: {james, nathan, matt}@dcs.warwick.ac.uk

***Abstract*—Conventions are often used in multi-agent systems to achieve coordination amongst agents without creating additional system requirements. Encouraging the emergence of robust conventions in a timely manner through the use of fixed strategy agents is one of the main methods of manipulating how conventions emerge. In this paper we demonstrate that fixed strategy agents can also be used to *destabilise* and remove established conventions. We examine the minimum level of intervention required to cause destabilisation, and explore the effect of different pricing mechanisms on the *cost* of interventions. We show that there is an inverse relationship between cost and the number of fixed strategy agents used. Finally, we investigate the effectiveness of placing fixed strategy agents by their cost, for different pricing mechanisms, as a mechanism for causing destabilisation. We show that doing so produces comparable results to placing by known metrics.**

***Keywords*—Conventions, Norms, Emergence, Cost, Destabilisation, Social Influence**

I. Introduction

Coordination is fundamental to multi-agent systems (MAS) and self-organisation as it enables systems to operate with increased efficiency and reduces the costs associated with mismatched actions. Coordinated actions are required since incompatible actions cause conflicts or incur costs. However, it is often impossible to constrain agents' actions ahead of time to ensure coordination. This can be due to a lack of *a priori* knowledge about clashing actions or the inability or unwillingness to dictate aspects of behaviour. This is of particular importance in self-* systems where there is no centralised control or where the range of possible actions makes pre-determination infeasible.

In response, many MAS rely on the emergence of conventions, in the form of the expected behaviour adopted by agents, with minimal prior involvement by system designers. As such, conventions allow coordinated actions to emerge through self-organisation, without needing to define behaviour beforehand. In particular, conventions have been shown to emerge given only agent rationality and the ability to learn from previous interactions. Understanding how they emerge and what system characteristics, such as network topology, might influence them is an area of active research [1], [2], [3], [4], [5].

Fixed strategy agents, that always choose the same action regardless of others' choices, have been shown to facilitate rapid convention emergence and to influence the adopted action. A small number of such agents, placed suitably, are able to influence a much larger population [2], [4], [6]. However, in realistic domains there is likely to be a cost associated with inserting a fixed strategy agent, or persuading an agent to act in a particular way, and it is desirable to minimise this cost.

It is useful to have the ability to replace existing conventions, as well as to establish new ones. Suboptimal conventions that have emerged, either due to restricted agent knowledge or a temporal quality of optimality, can be replaced with better conventions, increasing system efficiency. Understanding how these changes can be instigated will also allow the design of mechanisms that increase convention robustness, reducing the impact of outside influence.

This paper considers what is required 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 the established convention. Using this approach we show that a small proportion of IAs placed at targeted locations in the population for a sufficient length of time can destabilise an established convention, replacing it with another of our choosing. We also establish that the cost of these interventions varies inversely with the number of IAs used and that this effect is replicated across different pricing mechanisms. Finally we examine how the costs associated with agents may be used to inform where IAs are placed within the network. We show that, provided associated costs are a non-random indication of influence, placing by cost yields results similar to placing by network position metrics.

The remainder of this paper is structured as follows: in Section II we examine the related work on convention emergence and fixed strategy agents. Section III describes the model of convention emergence and metrics for characterising conventions used in this paper. The experimental setting is described in Section IV, and our results are presented in Section V. Finally, our conclusions are presented in Section VI.

II. Related Work

A *convention* is a form of socially-accepted rule regarding behaviour; there is no explicit punishment for going against the convention, nor any implicit benefit in the action represented by the convention over similar actions. The members of a convention expect others to act in a certain way, and deviation from the convention increases the likelihood of coordination problems and costs. A convention is "an equilibrium everyone expects in interactions that have more than one equilibrium" [7]. Conventions are able to emerge from local agent interactions [1], [3], [8], [9] and enhance coordination by placing *social constraints* on the actions that agents are likely to choose [10].

Conventions differ from *norms* (although the terms are sometimes used synonymously [4], [11]) as the latter imply an obligation to adhere with associated punishments for failing to perform the expected behaviour [12], [13], [14], [15]. Norms generally require additional system or agent abilities and incur an overhead to facilitate punishment for violation. In this paper we do not assume that agents have the capability to punish one another (or even to observe defections). Instead we focus on the use of conventions as a lightweight method of increasing coordination within the system.

Our only assumptions regarding agent behaviour are rationality and access to a (limited) memory of previous interactions. Convention emergence with these assumptions has been the focus of numerous studies [1], [4], [6], [9] and has been shown to allow rapid and robust convention emergence. Walker and Wooldridge [9] investigated convention emergence with few assumptions about the capabilities of the underlying agents. In their model agents select actions based on the observed choices of others, and global convention emergence is shown to be possible. Building on this, Sen and Airiau [4] explored social learning as a model for convention emergence, where agents receive a payoff from their interactions and use this to inform their learning. They show that convention emergence can occur when agents have no memory of interactions and are only able to observe direct interactions. However, their model is limited in that agents are not situated within a network topology and can interact with any member of the population, and the convention space has only two possible actions. In more realistic settings larger convention spaces and connecting network topologies are likely.

The effect of network topology and has been shown to have a significant effect on convention emergence [1], [3], [5], [16]. Recent work has investigated the effect of larger action spaces and has shown that a larger number of actions typically slows convergence [2], [6], [17].

The concept of utilising fixed strategy agents, those that always choose the same action regardless of others' choices, to influence convention emergence has also been explored. Sen and Airiau [4] show that a small number of such agents can cause a population to adopt the fixed strategy as a convention over other equally valid choices. This indicates that small numbers of agents are able to affect much larger populations.

Franks *et al.* [2], [18] investigated fixed strategy agents where interactions are constrained by a network topology with a large convention space. They found that topology affects the number of fixed strategy agents required to increase convergence speed. They also established that *where* such agents are placed is a key factor in how influential they are, with placement by metrics such as degree or eigenvector centrality being significantly more effective than random placement.

Previous work often assumes that there are no restrictions when placing fixed strategy agents into the network. We follow the assumption that such agents can be placed anywhere, but we assume that such an insertion has an associated *cost*. In real-world domains, inserting fixed strategy agents likely has a cost, and understanding how to minimise this cost is crucial. In this paper, we investigate the effect of the cost of insertion and its relation to the duration and efficacy of intervention.

Relatively little work has considered destabilising established conventions, with previous investigations of fixed strategy agents typically inserting them at the beginning of a simulation. We investigate inserting them when a convention has already become established with the aim of causing members of the dominant convention to change their adopted convention. Previous work has shown that destabilisation is possible [19], and in this paper we examine the minimum level of intervention required to cause destabilisation, and explore the effect of different pricing mechanisms and the effectiveness of placing fixed strategy agents by their cost.

Villatoro *et al.* [5], [8] develop techniques for destabilising meta-stable sub-conventions, which are stable subsets of the population adhering to secondary conventions. Meta-stable sub-conventions can slow adoption and prevent the emergence of an overall convention. Villatoro *et al.* identified particular topological substructures that are likely to cause meta-stable sub-conventions and target them in order to prevent them. However, their approach focusses on population segments whereas we examine destabilisation of the entire population. Moreover, in the work of Villatoro *et al.* the intention is to target meta-stable sub-conventions in order to facilitate better emergence of a primary convention, whilst we seek to destabilise an already established primary convention.

III. Convention Emergence Model

Convention emergence occurs as a result of agents in a population learning the best strategy over time. A population contains a set of agents, $Ag = \{1, ..., N\}$, who select from a set of possible strategies, $\Sigma = \{\sigma_1, \sigma_2, ..., \sigma_n\}$. Each timestep, every agent will choose one of its neighbours to perform an interaction with. Both choose an action from Σ and receive an individual payoff that is determined by the combination of actions. In this paper, the interaction and payoff are based on the n-action coordination game, such that agents receive a positive payoff if they select the same action and a negative payoff otherwise. The 2-action coordination game has frequently been used to explore convention emergence, but we generalise to the n-action coordination game to avoid restricting the number of possible conventions to a binary choice.

Each agent chooses the action that it believes will result in the highest payoff from knowledge of prior interactions. Agents also have the capability to explore the action space, such that with probability $p_{explore}$ agents will choose randomly from the available actions. In this regard we adopt the approach of Villatoro *et al.* [5], using a simplified Q-learning algorithm for both partners to update their strategies.

Additionally, agents are situated on a network topology that restricts their interactions to their neighbours. We consider small-world and scale-free topologies as these exhibit properties observed in real-world networks such as power law degree distributions and clustering. We also examine random topologies as a baseline.

A. Convention Metrics

To allow monitoring of convention establishment we need to formally state when a convention exists and identify the members of that convention. Previous work has adopted Kittock's criteria to define when conventions exist, such that a convention is considered to be established when 90% of the non-fixed-strategy agents, when not exploring, select the same action [3]. Whilst a useful method of defining established conventions, this criteria gives no way of examining *emerging* conventions, or of characterising a convention's decline if destabilisation occurs. Additionally, it presupposes the ability to examine agents' internals to establish when they are currently exploring, an ability that may not be possible in real-world domains. As such we utilise the metrics introduced in [19], which offer finer-grained measurement of conventions during their emergence, establishment and destabilisation. These are modified from the work of Walker and Wooldridge [9] and are defined as follows.

We begin by formalising what it means to say an agent chose a strategy:

$$\mathit{chose}_x(\sigma, t) \iff \exists i : i \in \mathit{par}_x(t) \wedge \mathit{self}_x(i, t) = \sigma$$

where $\mathit{self}_x(i, t)$ is the strategy chosen by agent x in interaction i in timestep t, and $\mathit{par}_x(t)$ is the set of interactions that x participated in during timestep t.

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

$$\mathit{chosen}(\sigma, t) = \{x | x \in Ag \wedge \mathit{chose}_x(\sigma, t)\}$$

We can now consider whether an agent is a member of a convention, and establish whether a particular convention exists. Due to exploration of the action space full adherence to a single strategy is unlikely. As such it is useful to quantify an agent's *adherence* to a strategy ϕ as the probability of that agent choosing ϕ in any potential interaction at time t:

$$adh(x, \phi, t) = P(\mathit{self}_x(i, t) = \phi \mid i \in \mathit{par}_x(t))$$

An exact measure of adherence is unlikely to be possible as it requires examination of agent internals for all but the simplest strategy selection methods. We can determine an *estimate* of adherence by examining an agent's interaction history, 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:

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

That is, a given strategy is considered to be a convention at time t if there is at least one agent using that strategy 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 a strategy that is also a convention and σ to denote a strategy that may or may not be a convention, allowing us to distinguish strategies 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 σ to be the mean adherence across all agents that chose σ in a timestep:

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

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 [9] note that strategy change has a cost and so the number of strategy changes can be expected to be minimised.

Using the average adherence, we define a convention as *established* if the average adherence is greater than the *convention establishment threshold* β:

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

Finally, we can define the extent to which agents are part of a convention. Agents are *members* of a convention if they currently have an adherence to it greater than or equal to β:

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

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

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

The sizes of the convention membership sets allow us to monitor the emergence, growth and destabilisation of conventions without having access to agent internals. We are also able to distinguish between agents who have chosen a strategy at random and those who are members of the convention.

B. Intervention Agents

As discussed in Section II, fixed strategy agents have been shown to affect convention emergence when placed in a population at the start of a simulation. In contrast to previous work we examine the effect of their introduction once a convention has emerged. We call such agents *Intervention Agents* (IAs) and, expanding on the work of Franks *et al.* [2], [18], we introduce IAs as replacements for agents within the primary convention (the convention with the highest membership) with the aim of destabilisation. The length of time these agents are left within the system is varied to explore the level of intervention needed to cause permanent change.

The strategy used by IAs depends on the aim of destabilisation. Where IAs use a specified convention the aim is to *promote* it whilst *demoting* the primary. Alternatively, the primary convention can be demoted whilst not explicitly specifying a replacement to promote, instead allowing a new convention to organically emerge. In this paper we focus on promoting the second most adopted convention and demoting the primary. Destabilisation without specifying an alternative convention is beyond the scope of this paper, and has been considered elsewhere [19].

IV. Experimental Setup

Our experimental setup is based on that presented used by Marchant *et al.* [19], in which a population of 1000 agents use Q-learning in the 10-action coordination game. The learning and exploration rate are both set to 0.25. Unless stated otherwise, all simulations are averaged over 30 runs.

An interaction window of $\lambda = 30$ is used for adherence approximation calculations. The probability threshold for an action to be considered a convention, γ, is set to 0.5 to enable more actions to be considered as conventions.

The established convention threshold, β, is set to 0.9. However, since we do not assume knowledge of whether an agent is exploring, the threshold must be reduced to account for random exploration. Therefore, we use: $\beta = 0.9 \times (1 - (p_{explore}(N-1))/N))$, where N is the number of strategies, $p_{explore}$ is the exploration rate, and $(N-1)/N$ represents the ratio of random choices that are not the "best" strategy.

Interaction topologies were generated using the Java Universal Network/Graph Library (version 2.0.1)[1]. 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 of the topology [20]. The Kleinberg model was used to generate the small-world topologies with a lattice size of 10×100, clustering exponent $\alpha = 5$ and one long distance connection per node [21]. As a baseline we also generated random network topologies using the Erdös-Rényi algorithm. To ensure that the densities of the graphs were similar a connection probability of 0.006 was used.

Simulations were run for 5000 timesteps before IAs were introduced, as convention emergence and stabilisation occurred within this time in all topologies. At timestep 5000, IAs were introduced, replacing nodes within the primary convention as selected by the placement strategy. Unless otherwise stated, the placement strategy was to select nodes in descending order of degree. The strategy adopted by IAs is that of the secondary convention at timestep 5000, i.e. the convention with the second highest membership. If multiple conventions have the same membership, the one with the highest average adherence is selected.

If there are insufficient members of the primary convention for the required number of IAs then additional IAs are placed elsewhere in the population, according to the placement strategy. Note that this implies the primary convention is immediately destabilised, as all its members become IAs, but such settings are included for completeness.

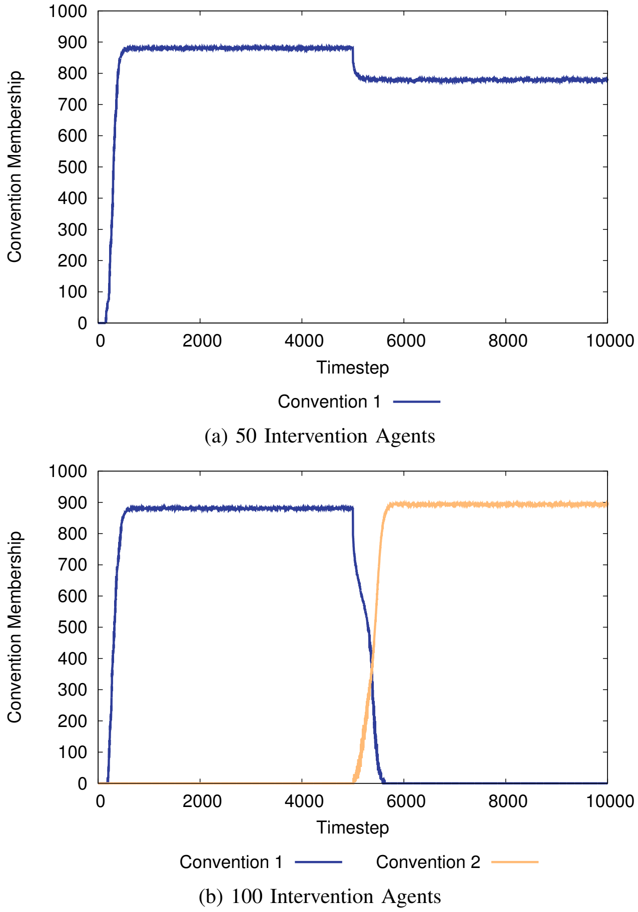
The IAs remain either until the end of the simulation or for a fixed number of timesteps, to investigate the duration required for destabilisation. When agents cease being IAs they again use Q-learning to choose actions (their learning continues during the time they are IAs). Unless otherwise stated, simulations ran for 10000 iterations in total, to give conventions after destabilisation enough time to emerge.

Each agent also has a non-negative *cost* associated with it. In order for an agent to be an IA this cost must be paid each timestep. As such, the cost of an intervention is simply the sum of the costs over all IAs for each timestep that the intervention occurs.

When considering the minimum cost of intervening we examine the idea of a *minimum intervention*, the minimum length of time that a given number of agents must remain in the system in order for destabilisation to occur. To quantify this we introduce a new measure: the crossover ratio χ_{co}. Given the membership level of the primary convention, $memb_{prim}$, and the membership level of the secondary convention, $memb_{sec}$, the crossover ratio is defined as:

$$\chi_{co} = \frac{memb_{sec}}{memb_{prim}}$$

The minimum intervention is the minimum amount of time that a given number of IAs must be introduced to cause χ_{co} to exceed some threshold, γ_{co}. In this paper we set $\gamma_{co} = 1.5$ such that the secondary convention must become 50% larger than the primary to be classed as destabilisation.



(a) 50 Intervention Agents

(b) 100 Intervention Agents

Fig. 1. The effect of IAs on convention membership in random graphs.

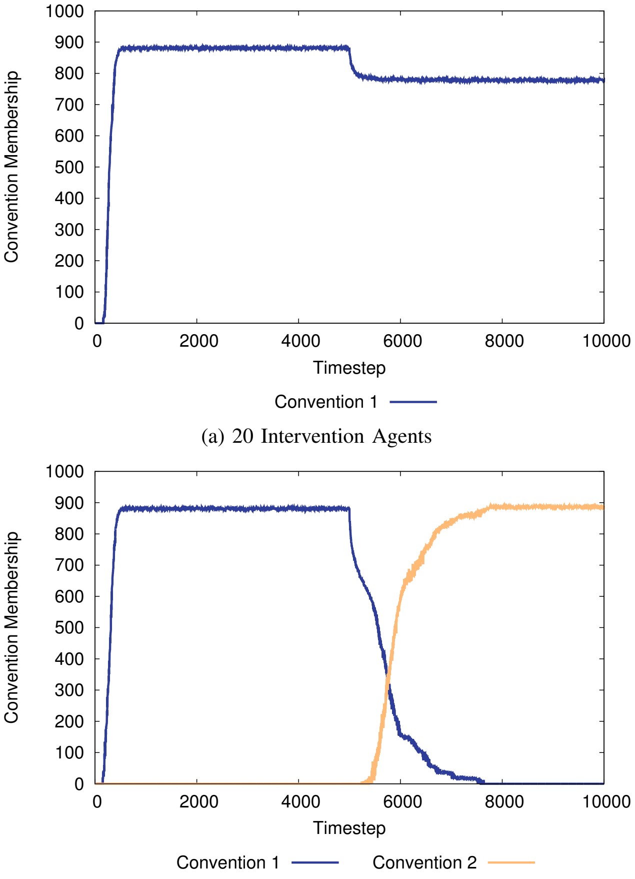
V. Results and Discussion

A. Number of fixed strategy agents

Our initial experiments seek to show that destabilisation is possible and establish the minimum number of IAs required. We begin by considering the setting where IAs remain in the system indefinitely after introduction, to remove the length of the intervention as a factor. Our initial experiments used both random and degree-based placement of IAs within the primary convention. Our results concur with the findings of Franks *et al.* [2], [18] and others that random placement of fixed strategy agents, whilst having the same overall effect, is inferior to placement by degree in terms of speeding up convention emergence. Therefore, in the remainder of this paper we focus on placement by highest degree.

Note that any conventions with zero or near-zero membership have been removed from the following figures for clarity,

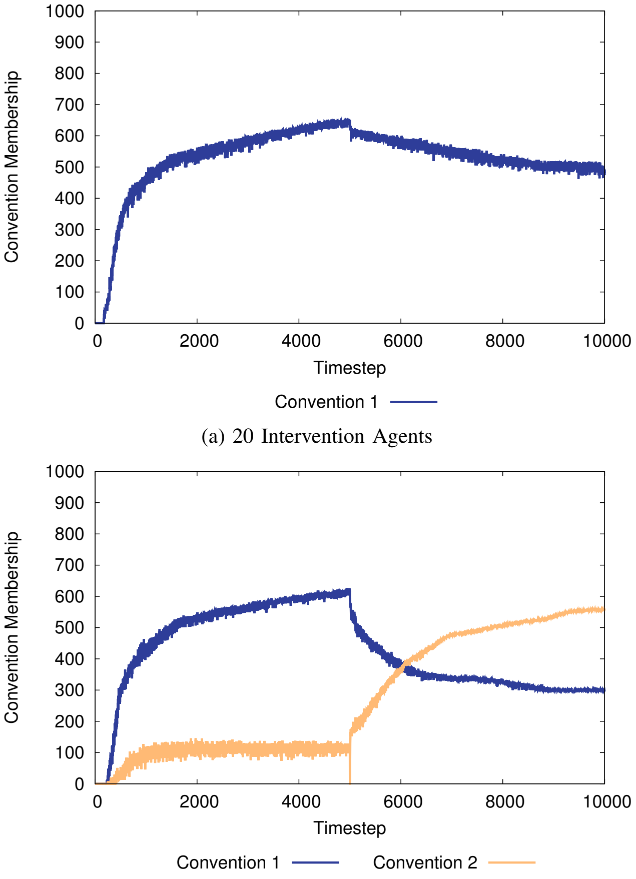
[1] http://jung.sourceforge.net/



(a) 20 Intervention Agents

(b) 40 Intervention Agents

Fig. 2. The effect of IAs on convention membership in scale-free graphs.



(a) 20 Intervention Agents

(b) 40 Intervention Agents

Fig. 3. The effect of IAs on convention membership in small-world graphs.

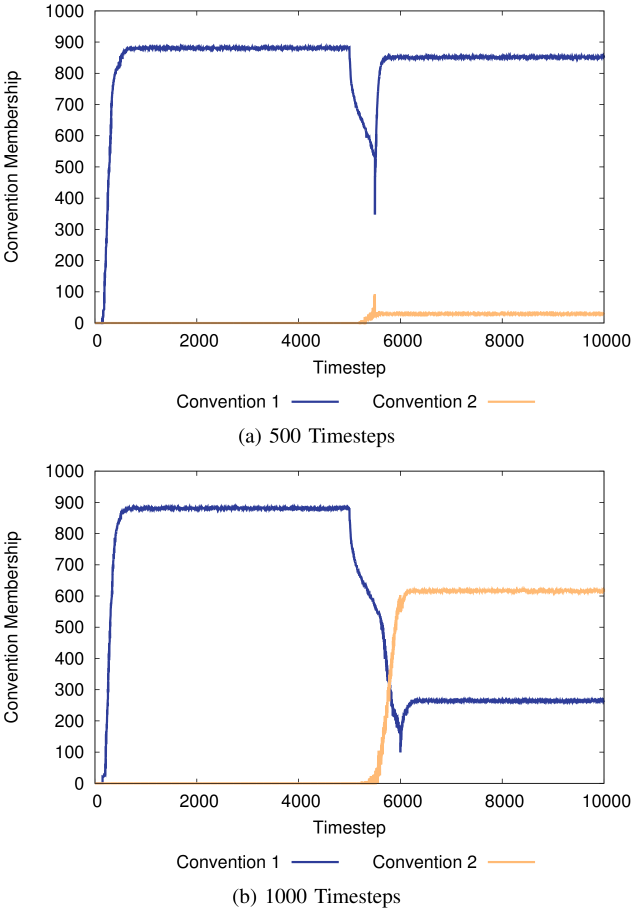
as they do not affect the emergence exhibited by the system. Conventions are labelled to indicate their relative rankings at timestep 5000.

To establish a baseline we begin by considering random topologies. Figure 1 shows the effect on convention membership of adding IAs to randomly generated graphs. As can be seen in Figure 1a, the addition of 50 IAs causes a drop in the membership of the primary convention after timestep 5000. The size of this drop is larger than that accounted for simply by the 50 agents who become IAs, indicating that the IAs are successful in changing the strategies of agents around them. However, the convention soon stabilises at a new level and the influence of the IAs ceases to spread. The secondary convention never becomes established, meaning that the average adherence to the strategy was too low, and those persuaded to move away from the primary convention did not become strong adherents to the secondary. In comparison, Figure 1b shows that insertion of 100 IAs causes the entire membership of the primary convention to switch, within only 1000 timesteps. These results show that there is a minimum number of IAs required to induce destabilisation, although fewer IAs than this minimum still cause the primary convention to stabilise at a lower level.

With a baseline established we now examine the effect on topologies that better represent the features found in real-world networks: scale-free and small-world. Figure 2 shows the results for scale-free topologies. The number of IAs required to exhibit the same behaviour as before is far fewer. In particular, destabilisation occurs with 40 IAs instead of 100. Fewer agents than this fail to cause destabilisation, achieving a drop in membership in the primary convention by influencing their local neighbourhood without propagating this further. This indicates that there is a topology-specific minimum number of IAs needed. The changeover in Figure 2b takes longer than in Figure 1b, but with larger numbers of IAs the speed of the transition increased. Thus, additional agents beyond the minimum speed up destabilisation.

The results for small-world topologies (shown in Figure 3) show similar behaviour to that presented above but there are some distinctions to highlight. Firstly, the overall level of membership is lower than in scale-free and random graphs and, secondly, changes take effect more gradually. Franks *et al.* [2] have observed similar differences between scale-free and small-world topologies and we hypothesise that the clustered nature of small-world networks is responsible for these effects. However, as in other topologies, a minimum number of agents is required to cause a destabilisation to occur.

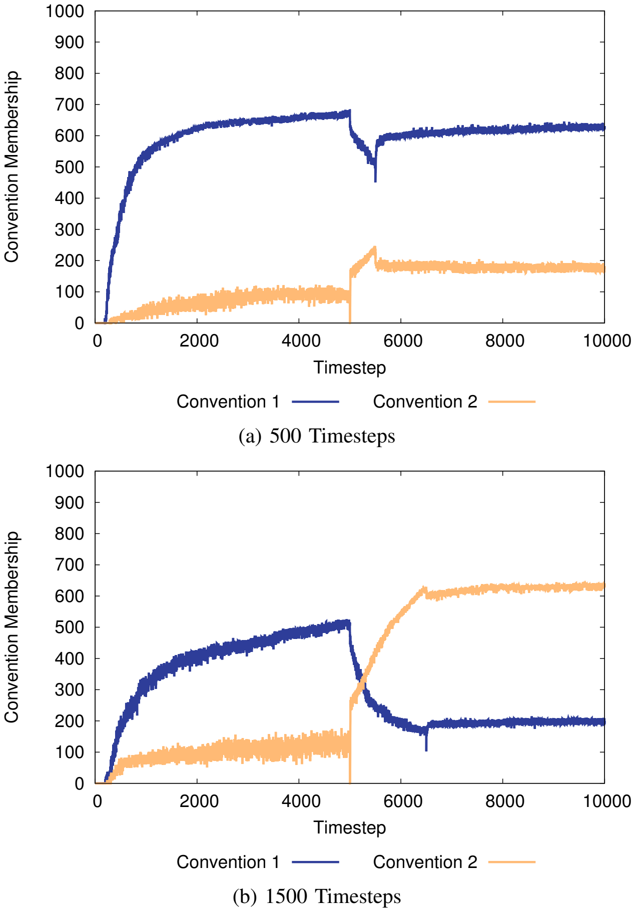
Within all three topologies there is some topology-specific minimum number of IAs that must be placed within the network in order for destabilisation to occur. Fewer IAs than this minimum allow the primary convention to stabilise at a lower level whilst additional IAs will increase the speed of



(a) 500 Timesteps

(b) 1000 Timesteps

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



(a) 500 Timesteps

(b) 1500 Timesteps

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

destabilisation.

B. Length of Intervention

Having established the minimum number of agents required to cause destabilisation we now examine the minimum duration that is required. Due to the similarities in behaviour between random and scale-free topologies, only scale-free and small-world networks are considered in the remainder of this paper.

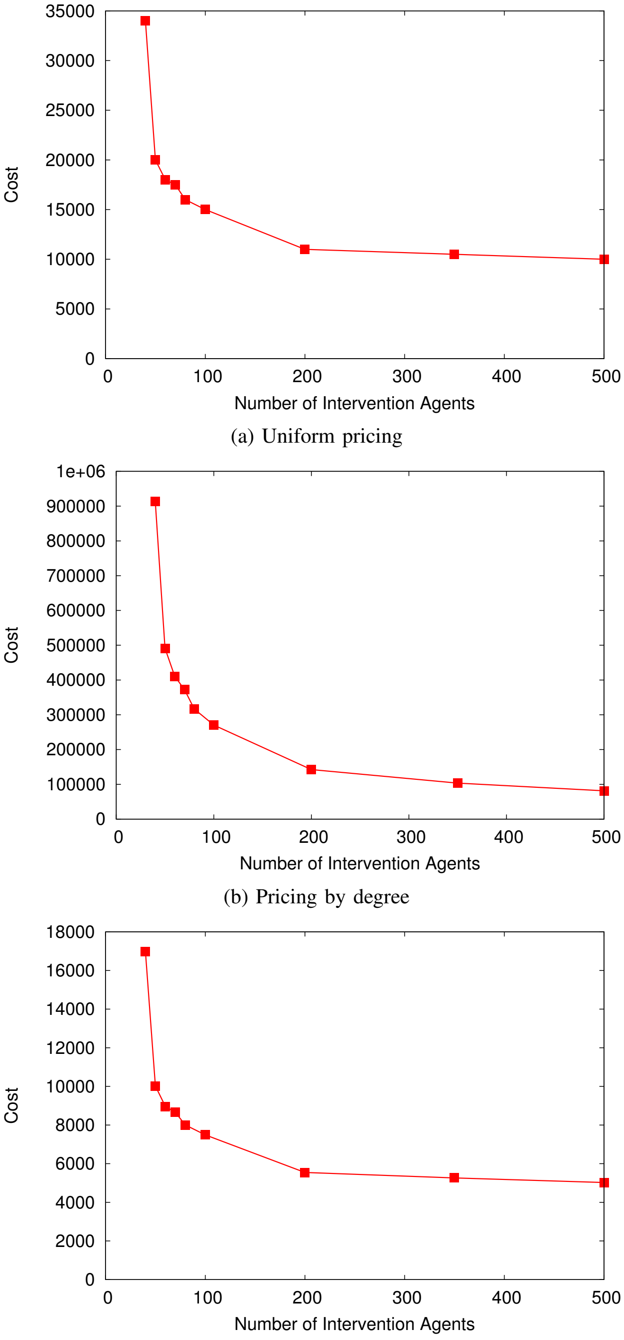
Whilst the previous simulations for scale-free networks allowed IAs to remain indefinitely, we now include them for finite time. This examines the ability of the primary convention to recover from temporary interventions. Figure 4 shows the effect of including 40 IAs for various durations. Figure 4a shows the behaviour when the IAs are inserted for 500 timesteps and then removed. Destabilisation begins to occur but, when the IAs are removed, the primary convention rapidly reclaims those agents who had changed. However, not all of them are reclaimed and the secondary convention has a small but notable membership after the intervention. This indicates that there is a minimum duration that IAs must be present to prevent the primary convention from reclaiming a significant proportion of the population. This is supported by Figure 4b where the IAs are present for twice as long and cause destabilisation to occur (though not fully). In this instance the primary convention reclaims some agents when the IAs are removed but stabilises at a far lower level than before the intervention.

Corresponding results for small-world topologies are presented in Figure 5. As before, the overall behaviour is the same, although there are some differences in small-world networks (namely the presence of a secondary convention before the intervention). The required IA duration was 1500 timesteps, which is significantly longer and is due to the more gradual adoption of change in small-world topologies.

Hence there is both a minimum number of IAs and a minimum length of time that they must be present in order for them to induce destabilisation. That is, there is a *minimum intervention* within each topology that must be met for conventions to undergo lasting change.

C. Cost of Intervention

To examine how the cost of destabilisation relates to the number of IAs used, we calculated the cost of the minimum intervention for various numbers of IAs. In order to find the minimum intervention for a given number of IAs the length of time that the IAs were inserted into the population was increased in steps of 50, starting from 0. For larger numbers of IAs (≥ 200) the length was increased in steps of 5, to add finer granularity to the minimum intervention approximation.

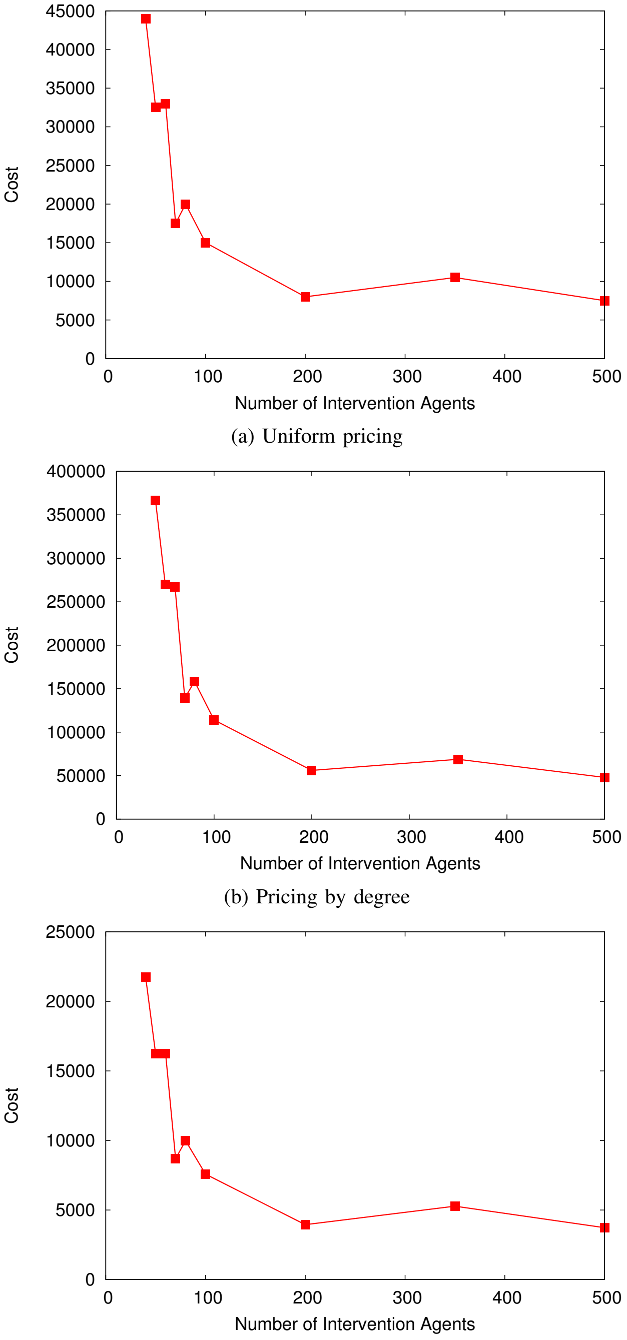


(a) Uniform pricing

(b) Pricing by degree

(c) Random pricing

Fig. 6. Number of IAs vs. the minimum cost to cause destabilisation for scale-free topologies. Placement is by degree and the pricing mechanism used is labelled on the graphs.



(a) Uniform pricing

(b) Pricing by degree

(c) Random pricing

Fig. 7. Number of IAs vs. the minimum cost to cause destabilisation for small-world topologies. Placement is by degree and the pricing mechanism used is labelled on the graphs.

The minimum intervention is defined as the smallest insertion time such that the crossover ratio of the averaged runs was greater than $\gamma_{co} = 1.5$. Whilst this is not the true minimum it gives an approximation which is sufficient for our calculations.

Initial experiments used a uniform price for all IAs, with each IA costing one unit each timestep. The number of IAs was varied from 40 (the minimum number needed to induce destabilisation in both topologies) up to 500. The latter is an unrealistically large proportion in most real-world domains, representing half of the population, but is included for completeness.

Figures 6a and 7a show the results for scale-free and small-world topologies[2] respectively. The cost of minimum intervention for both topologies decreases as the number of IAs increases, following an inverse relationship. Whilst the

[2] The artefacts present in all of the small-world results are due to how the minimum length was calculated. 80 and 90 IAs both had the same minimum length to the nearest 50 for instance.

cost per timestep increases, due to more IAs, the amount of time needed before destabilisation occurs decreases at a faster rate and hence the overall cost falls.

While both topologies exhibit this behaviour, the cost of minimum intervention in small-world topologies is generally higher than that in scale-free topologies, particularly with smaller numbers of IAs. This is again due to convention adoption in small-world topologies occurring at a slower rate than in scale-free networks and hence the IAs must remain for the same effect.

Uniform cost amongst agents is unlikely in real-world domains and so we consider additional pricing mechanisms. Figures 6b and 7b show results for the cost of minimum intervention where agents are priced directly based on their degree. Figures 6c and 7c present results where IAs have a cost selected at random from between 0 and 1. Note that in this set of results IA placement is by degree.

Whilst the scale of the graphs for these results vary, the relationship for each pricing mechanism is similar, with decreasing costs and diminishing returns as the number of IAs increases. This indicates that, regardless of how IA costs are calculated, it is cheaper to place as many IAs as possible into the system at high-degree locations. However the effect that this will have becomes substantially reduced after around 10% of the population (100 IAs).

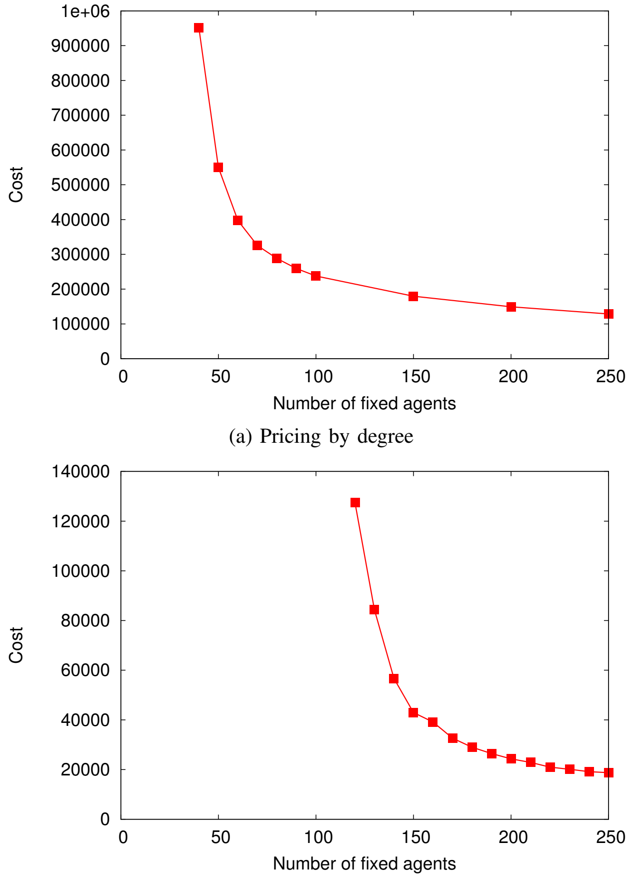
D. Cost-based Placement

In the above experiments we assume that information about the topology and agents' characteristics, such as degree, are available. We now consider the situation where information such as the degree of agents is hidden, and all that is known is an *advertised cost* which may or may not be a good indication of an agent's influence.

Figures 6c and 7c correspond to the case where the cost gives no indication of influence. However, in these simulations the degree of agents was still available and used for placement decisions. In the following experiments, IAs are placed at *high-cost* locations, rather than assuming knowledge of degree.

Our previous experiments also assumed that multiple simulations could be performed ahead of time, slowly increasing the duration of intervention to find the minimum effective duration. In real-world settings this is impractical and instead an intervention must be monitored in real-time to establish whether destabilisation has occurred and the IAs can be removed. In the following experiments we use moving averages (with a window size of 30 timesteps) to calculate the current χ_{co} within a simulation. When this exceeds γ_{co} we consider destabilisation to occur and the IAs are removed and the simulation terminated. The cost up to this point represents the cost of a minimum intervention. If this condition is not met by timestep 10000 then the run is deemed unlikely to destabilise and is marked as invalid. For the minimum interventions to be considered representative (rather than occurring by chance), $2/3$ of the runs must be valid. If this condition is met then the average minimum cost over the valid runs is calculated.

We begin by considering the effect of pricing (and hence also placing) agents by degree in scale-free networks, with the results shown in Figure 8a. Whilst this setting has the same placement and cost mechanism as presented in Figure 6b, the method of calculating minimum interventions is distinct. That the two sets of results are similar indicates that the two methods are equally valid for determining the minimum intervention, and provides a basis for further pricing mechanisms.



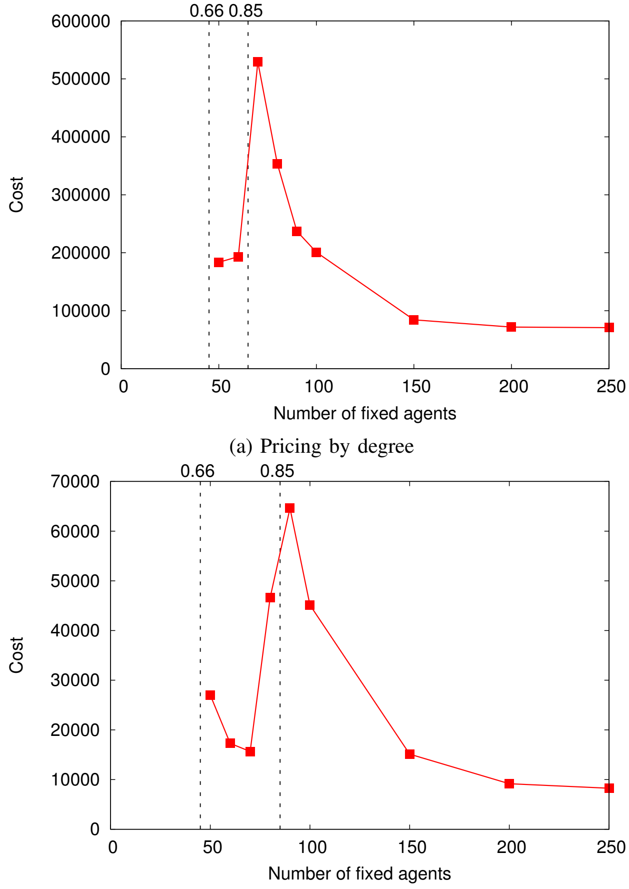
(a) Pricing by degree

(b) Random pricing

Fig. 8. Number of IAs vs. the minimum cost to cause destabilisation for scale-free topologies. Agents are placed at high cost locations.

In order to change from degree-based placement we examine the process of pricing and placing randomly. Figure 8b presents these results. Whilst the relationship between cost and the number of IAs remains, a larger number of IAs is needed to give sufficient valid runs. This is to be expected, as similar results regarding random placement were found above and in previous work. Importantly, even when placing randomly, we see an inverse relationship regarding the number of IAs and cost.

These pricing mechanisms were also examined for small-world topologies. Figure 9a shows the effect of placing and pricing by degree for these topologies. Unlike in scale-free networks, the behaviour for low numbers of IAs differs from that found previously. However, when 80 or more IAs are introduced the relationship matches that of Figure 7b, which is expected. Further experimentation revealed that this is due to the variation between different runs on small-world topologies being greater than in scale-free topologies. The ratio of valid runs for degree-based cost placement in small-world topologies is shown in Figure 10. As the number of agents increases the



(a) Pricing by degree

(b) Random pricing

Fig. 9. Number of IAs vs. the minimum cost to cause destabilisation for small-world topologies. Agents are placed at high cost locations.

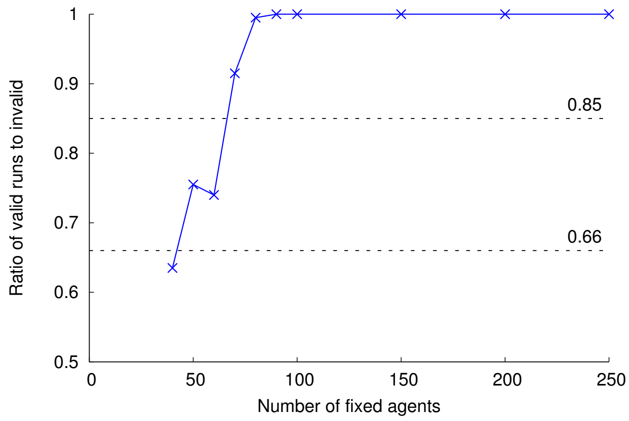
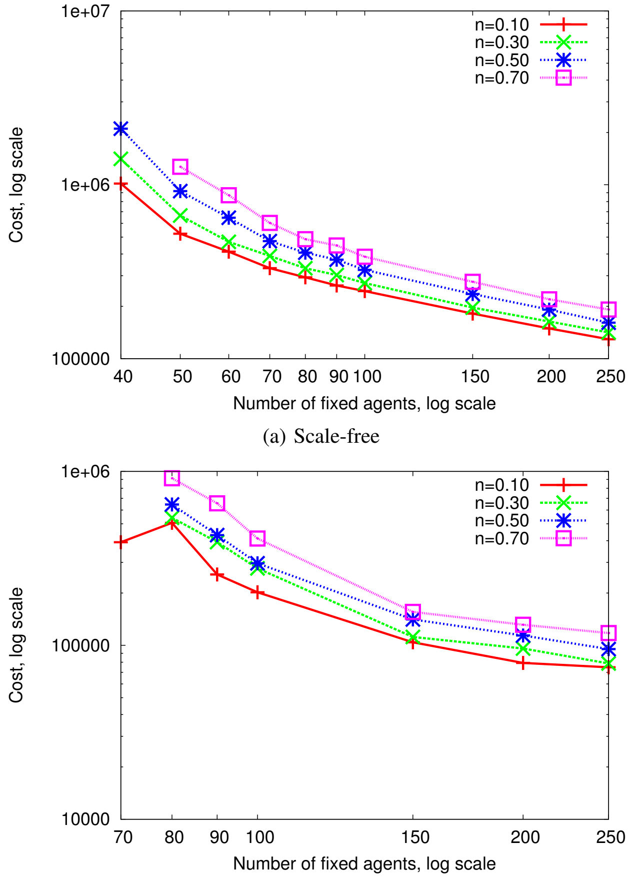


Fig. 10. Number of IAs vs. valid run ratio in small-world networks with pricing and placement by degree.

proportion of valid runs increases. In comparison, scale-free topologies were found to exhibit a binary valid run ratio, such that either all runs were valid or none were. As such, the $2/3$ valid runs threshold correctly captures when destabilisation is consistently occurring in scale-free topologies. For small-world topologies the gradual change means that runs that contain destabilisation but are not representative will be counted, as indicated by the lower dotted line in Figure 10. Setting the valid run threshold higher (0.85 is also shown) means that valid runs will be better indicators of minimum interventions. The effect of alternative valid run thresholds is shown in Figure 9. This gives an important insight into the nature of small-world topologies, namely that individual simulations are less predictable for lower numbers of IAs than their scale-free equivalents.



(a) Scale-free

(b) Small-world

Fig. 11. Number of IAs vs. the minimum cost to cause destabilisation for scale-free topologies. The pricing mechanism is degree with additional noise with IAs placed at high cost locations. Noise is varied from 0.1 to 0.7 in steps of 0.2.

Finally we examine the situation where the advertised cost of an agent is an imperfect indication of their degree (and hence influence). This pricing mechanism is useful in many real-world domains where agents may be asked to estimate their own influence or where topology information is unreliable. This is modelled by selecting each agent's advertised cost from a Gaussian distribution:

$$cost(v) = \mathcal{N}\left(deg(v), (deg(v) \times noise\_level)^2\right)$$

That is, the cost is equal to the degree plus Gaussian noise with a standard deviation equal to some fraction of the degree.

Results for this setting are shown in Figure 11. For both scale-free and small-world topologies the noise level, n, was varied from 0.1 to 0.7, and the valid run ratio threshold was set to 0.85 to remove the artefacts present in small-world graphs.

The effect in both topologies of increasing noise is to increase the overall cost that is needed to cause destabilisation. The results are shown on a log-log scale to more easily distinguish this. However, even with 70% noise being applied, the relationship between cost and number of IAs remains the same. As long as the cost is known to be a function of degree, rather than truly random, it is beneficial to base placement decisions on this information even if it is substantially noisy.

Amongst all pricing mechanisms the same inverse relationship between number of IAs and overall cost remains. However, the number of IAs required to consistently cause destabilisation is affected by this mechanism. Hence the best strategy is still to insert as many IAs as possible, using advertised cost if no other metrics are available.

VI. Conclusions

We have shown that it is possible to cause destabilisation of existing conventions by the insertion of a small proportion of fixed strategy *Intervention Agents* into the population at key locations. By setting the strategy of these agents to that of the second largest convention we have shown that the primary convention can be destabilised and replaced with the secondary. In scale-free and small-world topologies we found that 40 IAs in a population of 1000 were sufficient to cause this, whilst in random topologies 100 IAs were needed. Fewer IAs than this were shown to cause a fall in the membership of the primary convention in each topology, but not enough to make the secondary convention dominant.

We have also shown that temporarily inserting IAs can also cause destabilisation, and that there exists a minimum length of time that they must be present in order to cause this. Removing IAs prior to this minimum duration will cause the primary convention to return to near previous levels. We found that the minimum length of time required was smaller in scale-free topologies than small-world topologies.

Next we considered the cost of these interventions, and show that, independent of whether cost is random, uniform, or linked to degree, the cost of minimum intervention is inversely related to the number of IAs. However, the relationship is one of diminishing returns. As such, placing as many IAs as possible into the system is beneficial but the additional effect generated reduces substantially after 10% of the population.

Finally, we explored the effect of placing IAs by cost and monitoring destabilisation in real-time. The same relationship between number of IAs and cost was found to hold regardless of pricing/placement mechanism although higher numbers of IAs may be needed to sufficiently guarantee destabilisation. We also found that small-world topologies vary in this respect more between simulations than scale-free networks. The effect of noise on the degree-based pricing mechanism was also considered. It was found, for both topologies, that the effect of noise was to increase the overall cost of minimum interventions but to not affect the relationship between cost, the number of IAs, and the duration of minimum interventions. We conclude from this that placing by advertised cost would offer reasonable results, assuming non-random pricing.

Overall we have shown that destabilisation and replacement of an established convention is possible and that minimum criteria exist in order to cause this. We have also presented a number of ways of evaluating how much an intervention might cost using various pricing methods and demonstrated the relationship between number of IAs and cost.

References

[1] J. Delgado, J. M. Pujol, and R. Sangüesa, "Emergence of coordination in scale-free networks," *Web Intelli. and Agent Sys.*, vol. 1, no. 2, pp. 131–138, 2003.

[2] H. Franks, N. Griffiths, and A. Jhumka, "Manipulating convention emergence using influencer agents," *Autonomous Agents and Multi-Agent Systems*, vol. 26, no. 3, pp. 315–353, 2013.

[3] J. Kittock, "Emergent conventions and the structure of multi-agent systems," in *Lectures in Complex Systems: the Proc. of the 1993 Complex Systems Summer School.* Addison-Wesley, 1995, pp. 507–521.

[4] S. Sen and S. Airiau, "Emergence of norms through social learning," in *Proceedings of the 20th International Joint Conference on AI.* Morgan Kaufmann Publishers Inc., 2007, pp. 1507–1512.

[5] D. Villatoro, S. Sen, and J. Sabater-Mir, "Topology and memory effect on convention emergence," in *Proc. of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology.* IEEE Computer Society, 2009, pp. 233–240.

[6] N. Griffiths and S. S. Anand, "The impact of social placement of non-learning agents on convention emergence," in *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems.* International Foundation for Autonomous Agents and Multiagent Systems, 2012, pp. 1367–1368.

[7] H. P. Young, "The economics of convention," *The Journal of Econ. Perspectives*, vol. 10, no. 2, pp. 105–122, 1996.

[8] D. Villatoro, J. Sabater-Mir, and S. Sen, "Social instruments for robust convention emergence," in *Proc. of the 22nd International Joint Conference on AI.* AAAI Press, 2011, pp. 420–425.

[9] A. Walker and M. Wooldridge, "Understanding the emergence of conventions in multi-agent systems." in *International Conference on Multi-Agent Systems.* MIT Press, 1995, pp. 384–389.

[10] Y. Shoham and M. Tennenholtz, "On the emergence of social conventions: modeling, analysis, and simulations," *Artificial Intelligence*, vol. 94, no. 1–2, pp. 139–166, 1997.

[11] P. Mukherjee, S. Sen, and S. Airiau, "Norm emergence with biased agents," *International Journal of Agent Technologies and Systems*, vol. 1, no. 2, pp. 71–84, 2009.

[12] R. Axelrod, "An evolutionary approach to norms," *American Polit. Sci. Rev.*, vol. 80, pp. 1095–1111, 1986.

[13] M. Kandori, "Social norms and community enforcement," *The Rev. of Econ. Studies*, vol. 59, no. 1, pp. 63–80, 1992.

[14] C. Bicchieri, R. C. Jeffrey, and B. Skyrms, *The Dynamics of Norms.* Cambridge University Press, 1997.

[15] T. B. R. Savarimuthu, R. Arulanandam, and M. Purvis, "Aspects of active norm learning and the effect of lying on norm emergence in agent societies," in *Agents in Principle, Agents in Practice*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2011, vol. 7047, pp. 36–50.

[16] J. Delgado, "Emergence of social conventions in complex networks," *Artificial Intelligence*, vol. 141, no. 1–2, pp. 171–185, 2002.

[17] N. Salazar, J. A. Rodriguez-Aguilar, and J. L. Arcos, "Robust coordination in large convention spaces," *AI Commun.*, vol. 23, no. 4, pp. 357–372, 2010.

[18] H. Franks, N. Griffiths, and S. S. Anand, "Learning agent influence in MAS with complex social networks," *Autonomous Agents and Multi-Agent Systems*, pp. 1–31, 2013.

[19] J. Marchant, N. Griffiths, M. Leeke, and H. Franks, "Destabilising conventions using temporary interventions," in *Proceedings of the 17th International Workshop on Coordination, Organizations, Institutions and Norms*, 2014.

[20] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Rev. of Mod. Phys.*, vol. 74, no. 1, pp. 47–95, 2002.

[21] J. Kleinberg, "Navigation in a small world," *Nature*, vol. 406, no. 6798, pp. 845–845, 2000.