Norm Emergence through Dynamic Policy Adaptation in Scale Free Networks

Samhar Mahmoud[1], Nathan Griffiths[2], Jeroen Keppens[1], and Michael Luck[1]

[1] Department of Informatics, King's College London, London WC2R 2LS, UK.
samhar.mahmoud@kcl.ac.uk
[2] Department of Computer Science, University of Warwick, Coventry CV4 7AL, UK.

Abstract. As has been stated elsewhere, norms are a valuable means of establishing coherent cooperative behaviour in decentralised systems in which there is no central authority. Axelrod's seminal model of norm establishment in populations of self-interested individuals provides some insight into the mechanisms needed to support this through the use of *metanorms*, but considers only limited scenarios and domains. While further developments of Axelrod's model have addressed some of the limitations, in particular in considering its application to different topological structures, this too has been limited in not offering an effective means of bringing about norm compliance in scale-free networks, due to the problematic effects of *hubs*. This paper offers a solution, first by adjusting the model to more appropriately reflect the characteristics of the problem, and second by offering a new dynamic policy adaptation approach to learning the right behaviour. Experimental results demonstrate that this dynamic policy adaptation overcomes the difficulties posed by asymmetric distribution of links in scale-free networks, leading to an absence of norm violation, and instead norm emergence.

1 Introduction

Norms are an effective means of governing the behaviours of different members of decentralised open systems, such as P2P file-sharing systems in which cooperation between members maintains benefits for all. However, individuals often take benefits without contributing to the common good, the *free riding* phenomenon [1] by which some download files from others without uploading in return. In decentralised systems, the absence of a central authority means that there is no consequence for such behaviours. Many researchers ([4, 6, 7, 12, 14, 16]) have proposed norms as a means of regulating agent behaviour but, as shown by Axelrod [2], norms alone may not lead to desired outcomes. In consequence, Axelrod proposed *metanorms* as a means of ensuring not that norms are complied with, but that they are enforced. He showed that metanorms are effective in fully-connected networks, but did not consider other kinds of topology.

Some work has already been undertaken on examining the impact of different topologies on norm establishment. For example, Savarimuthu et al. [11] consider the *ultimatum game* in the context of a role model that provides advice on whether to change norms in order to enhance performance, and provide experimental results for random and scale-free networks. Delgado et al. [5] study norm emergence in coordination

games in scale-free networks, and Sen et al. [13] similarly examine rings and scale-free networks in a related context. Additionally, Villatoro et al. [15] explore norm emergence within lattices and scale-free networks. While these efforts provide valuable and useful results, the context of application has tended to be limited, with only two agents involved in a single interaction, rather than a larger population. This simplifies the problem compared to those in which *multiple* agents are involved in a single interaction can impact on norm establishment.

Rather than adopting a fundamentally different model, in this paper we examine the problem of norm establishment in Axelrod's original model but extended to address the issues arising in topological structures, and in particular scale-free networks, which cause two significant problems. First, Axelrod's model assumes a fully connected network, and is predicated on that for certain aspects, such as how one agent observes another's actions. In a variably connected structure, this part of the model is thus not meaningful and requires modification, causing some difficulties in establishing norms. Second, in scale-free networks, which contain both heavily connected nodes (*hubs*) and lightly-connected nodes (*outliers*), hubs strongly influence norm emergence since they are involved in observation of, and interaction with, so many others in the network. While the work of Galan et al. [8] addresses the first point, applying Axelrod's model to other networks, the approach requires inappropriate access to the strategy of others [10].

In response, this paper provides two key contributions: it addresses a weakness in a previous technique for lattices and small-worlds to be consistent with the requirements of agent autonomy, and it provides a dynamic policy adaptation mechanism that leads to norm emergence in scale-free networks for which prior efforts have not succeeded. The paper begins with a brief description of the *metanorm model*. Section 3 then considers the problems that arise from the use of scale-free networks, and the adaptation of the model to cope with their characteristics. Section 4 introduces our solution for achieving norm emergence in this context and, finally, Section 5 concludes.

2 The Metanorms Model

Inspired by Axelrod's model [2], our simulation focusses only on the essential features of the problem. In the simulation, the agents play a game iteratively; in each iteration, they make a number of binary decisions. First, each agent decides whether to comply with the norm or to defect. Defection brings a reward for the defecting agent, and a penalty to all other agents, but each defector risks being observed by the other agents and punished as a result. These other agents thus decide whether to punish agents that were observed defecting, with a low penalty for the punisher and a high penalty for the punished agent. Agents that do not punish those observed defecting risk being observed themselves, and potentially incur metapunishment. Thus, finally, each agent decides whether to metapunish agents observed to spare defecting agents. Again, metapunishment comes at a high penalty for the punished agent and a low penalty for the punisher.

The behaviour of agents in each round of the game is random, but governed by three variables: the probability of being seen, *boldness*, and *vengefulness*. In each round, agents have a fixed number of opportunities to defect, each of which has a randomly selected probability of a defection being seen. Then, if an agent's *boldness* exceeds the

Algorithm 1 The Simulation Loop

```
1. for each round do
2.     interact()
3.     learn()
```

probability of a defection being seen, the agent defects. *Vengefulness* is the probability that an agent punishes or metapunishes another agent. Thus, the boldness and vengefulness of an agent are said to comprise that agent's *strategy* (or *policy* which we use when referring more specifically to agent actions). After several rounds of the game, each agent's rewards and penalties are tallied, and successful and unsuccessful strategies are identified. By comparing themselves to other agents on this basis, the strategies of poorly performing agents are revised such that features of successful strategies are more likely to be retained than those of unsuccessful ones. While successful strategies for agents increase their scores, our aim, in contrast, is to find strategies that will establish norms in the community: i.e. where agents tend to abide by the norm (due to low boldness) and tend to enforce and meta-enforce the norm (due to high vengefulness).

2.1 Evolution vs Learning

Axelrod's evolutionary approach works well as a model to identify the *fittest* strategies because all agents act in an identical context. With the introduction of a topology, each agent possesses a potentially unique context defined by the agents to which it is connected. As such, strategies are not freely interchangeable. An agent typically keeps its strategy private, but adapts it in response to perceived effectiveness. Mahmoud et al. addressed these concerns by means of a reinforcement learning mechanism [10]; since we base our work on this starting point, we briefly describe those algorithms before presenting our own contributions. As illustrated in Algorithm 1, the simulation consists of several rounds, comprising an *interaction phase* and a *learning phase*, both of which rely on the identification of different components in the makeup of the utility (or *score*), which we introduce first. In order to assess and improve performance, each agent has a measure of utility, represented by different *scores*: defection score (DS), punishment score (PS), punishment omission score (POS) and total score (TS). DS is the utility gained or lost from defecting, PS is the utility lost from punishing others, and POS is the utility lost from not punishing defections. Finally, TS is the sum of these scores, in addition to the loss of utility that is borne by agents as a result of others defecting.

The interaction phase is concerned with the interaction between agents. As shown in Algorithm 2, each agent has various defections opportunities (o), in which it decides to defect (to violate the norm) if its boldness is greater than the probability of its defection being seen (S_o). As a result, the agent's DS is increased by a temptation value T, while every neighbouring agent NB (all agents that are directly connected) is hurt, reflected through a negative value H being added to the agent's total score. However, each hurt agent can observe the defection and react to it by imposing a punishment with probability corresponding to the observing agent's vengefulness. If an agent is punished, then its DS is decreased by a value P, and the punishing agent's PS is decreased with enforcement cost E. If an observing agent does not punish then, in

Algorithm 2 interact()

```
1.  for each agent i do
2.    for each opportunity to defect o do
3.      if B_i > S_o then
4.        DS_i = DS_i + T
5.        for each agent j ∈ NB_i: j ≠ i do
6.          TS_j = TS_j + H
7.          if see(j,i,S_o) then {j sees i}
8.            if punish (j, i, V_j) then {j punishes i}
9.              DS_i = DS_i + P
10.             PS_j = PS_j + E
11.           else
12.             for each agent k ∈ NB_j : k ≠ i ∧ k ≠ j do
13.               if see(k,j,S_o) then
14.                 if punish (k, j, V_j) then
15.                   PS_k = PS_k + E
16.                   POS_j = POS_j + P
```

turn, neighbours that observe this can metapunish the agent, again with probability corresponding to vengefulness. This results in the metapunished agent's POS being decreased by P (and thus increased in magnitude), since the metapunishment is a result of not punishing the defector, while the metapunishing agent's PS is reduced by E.

In the learning phase (Algorithm 3), the various scores are used as a means of improving performance in each round. Agents change their policies for action in the direction that should result in better scores. Initially, TS is calculated by accumulating the various component scores, and this is then used to determine whether to modify its policy, by comparing TS with the average population score (since agents that perform well should not change). If an agent's defection score DS is positive then it increases boldness, and decreases it if negative. Conversely, vengefulness is increased if PS is better than POS, and decreases otherwise. These changes arise by adding or subtracting a learning rate δ. Moreover, to explore the policy space, an agent may completely change its boldness and vengefulness values, determined by an *exploration rate*, γ.

2.2 Metanorms, Lattices and Small Worlds

As Mahmoud et al. [9] demonstrate, applying this model to fully-connected networks, lattices and small worlds results in norm emergence with different levels of success corresponding to the characteristics of the topologies. More specifically, the model always results in a population of agents with low average boldness and varying degrees of high average vengefulness. However, both lattices and small worlds have the attribute that *neighbourhood size* determines the number of neighbours to which each agent must be connected, and this appears to be important for convergence to norm emergence, with larger neighbourhoods giving better vengefulness. Conversely, *population size* has no effect on lattices, but in small worlds a larger population decreases vengefulness.

Algorithm 3 learn()

```
1.  for each agent i do
2.    TS_i = TS_i + DS_i + PS_i + POS_i
3.    if TS_i < AvgS_{NB_i} then
4.      if explore(γ) then
5.        B_i = random()
6.        V_i = random()
7.      else
8.        if DS_i < 0 then
9.          B_i = B_i − δ
10.       else
11.         B_i = B_i + δ
12.       if PS_i < POS_i then
13.         V_i = V_i − δ
14.       else
15.         V_i = V_i + δ
```

3 Scale-free Networks

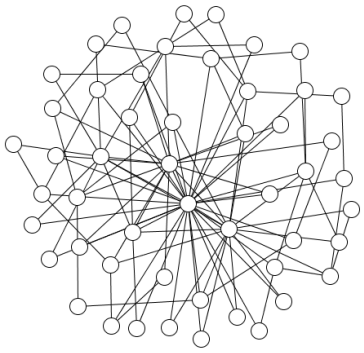


Fig. 1. Example of a Scale-free network

The topologies considered above are similar in that each agent has exactly the same number of connections, in contrast to scale-free networks [3], in which connections between nodes follow a power law distribution. Thus, some nodes have a vast number of connections, but the majority have very few connections, as illustrated in Figure 3. These properties of scale-free networks suggest an imbalance in connections. In turn, this has an impact on the results that can be obtained, due both to punishment and to enforcement costs, which dramatically modify the dynamics of the system. To investigate this, we ran 1000 experiments on a scale-free network with 1000 agents, five of which were *hubs* (having a large number of connections) and the others (which we call

outliers) with at least two connections to other agents in the population, and typically no more than four connections (according to Barabasi's algorithm [3]). Each experiment was run for 1000 rounds (or timesteps), and parameters for the experiments were as follows (and are the same for all subsequent experiments reported in this paper): $T = 3$, $E = -2$, $P = -9$, $H = -1$, $\delta = \frac{1}{7}$ and $\gamma = 0.01$. The results, shown in Figure 3, indicate that all runs end with both average boldness and average vengefulness, so that no norm is established. However, a detailed analysis of individual runs reveals that this is because there is no significant change to the average vengefulness and boldness, with both fluctuating around the average from the start of the run until the end.

By differentiating between hubs and outliers, some patterns are revealed, however. In particular, the model succeeds in lowering the boldness of hubs, but their vengefulness remains near average. Because hubs are connected to many other agents and are punished many times for a defection, boldness decreases. Conversely, they also punish many of these other agents for defecting, and consequently pay a very high cumulative enforcement cost that causes them to lower their vengefulness. In turn, this lower vengefulness causes them subsequently not to punish others and as a result to receive metapunishment from other hubs, leading to an increase in vengefulness again. Over time, this repeats, with vengefulness decreasing and then increasing back to the average, as shown in Figure 3(a) and 3(c). Note that for all results that are provided in this paper, we show both the long terms results of 1000 timesteps for completeness and the short term results of 100 timesteps for clarity.

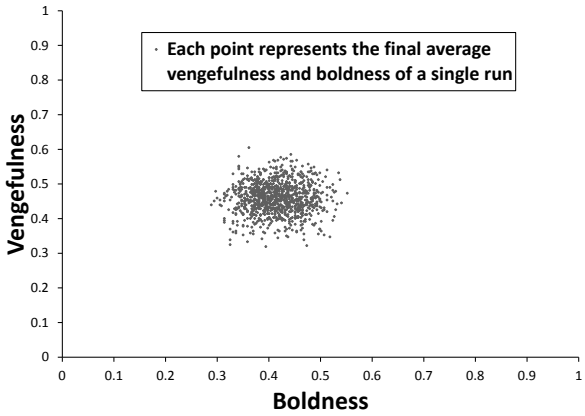
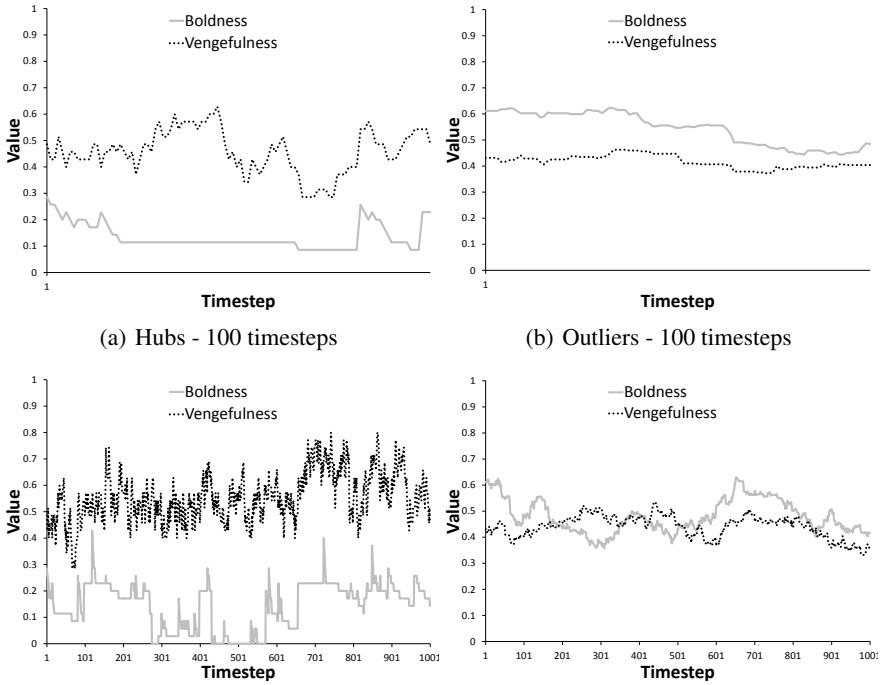


Fig. 2. Overall Result - 1000 Runs, 1000 Agents, 1000 Timesteps

For the remaining, *outlier*, agents, changes to boldness and vengefulness are indicative of overall boldness and vengefulness because they comprise the majority of the population. They are typically connected to one or more of the hubs, and while they too defect and punish, they do so much less frequently than the hubs to which they are connected. Thus, their scores are higher than the scores of the hubs; because those agents with higher scores do not learn from others (since there are no higher scoring others to

learn from), they do not change their strategies, and their boldness and vengefulness remains close to the average, as shown in Figures 3(b) and 3(d). These results demonstrate that Mahmoud et al.'s algorithm is not effective in scale-free networks. Importantly, as the burden of punishment falls largely on hubs rather than outliers, hubs perform worst in the population. To address this, we modify the learning technique so that it can cope with the nature of scale-free networks, as discussed next.



(a) Hubs - 100 timesteps (b) Outliers - 100 timesteps

(c) Hubs - 1000 timesteps (d) Outliers - 1000 timesteps

Fig. 3. Sample Run

3.1 Universal Learning

The algorithm proposed by Mahmoud et al. suffers from the limitation that it requires knowledge of the average score in the population in order for an agent to determine whether to modify its policies. However, since the aim of that work is to eliminate the unreasonable assumption of *omniscience*, by which agents are able to observe the private strategies of others, as well as observing all norm violations and punishments, it makes little sense to assume that agents have access to an average population score against which to compare themselves before deciding whether to modify their policies. For this reason, we consider an alternative approach, in which agents always modify

their policies to improve performance, regardless of the behaviour of others, and only in relation to their own score. This modification is simple, and involves removing line 3 of Algorithm 3 (we do not show the new algorithm due to the simplicity of the change and space constraints).

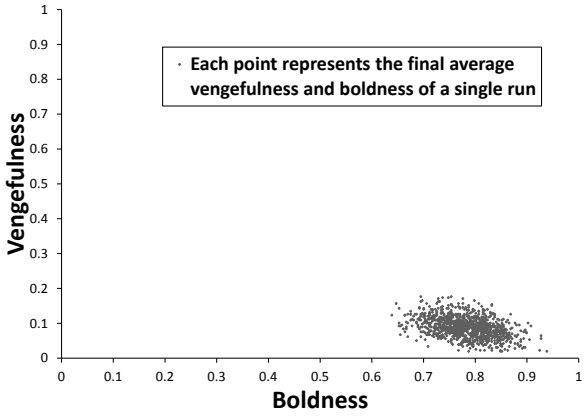
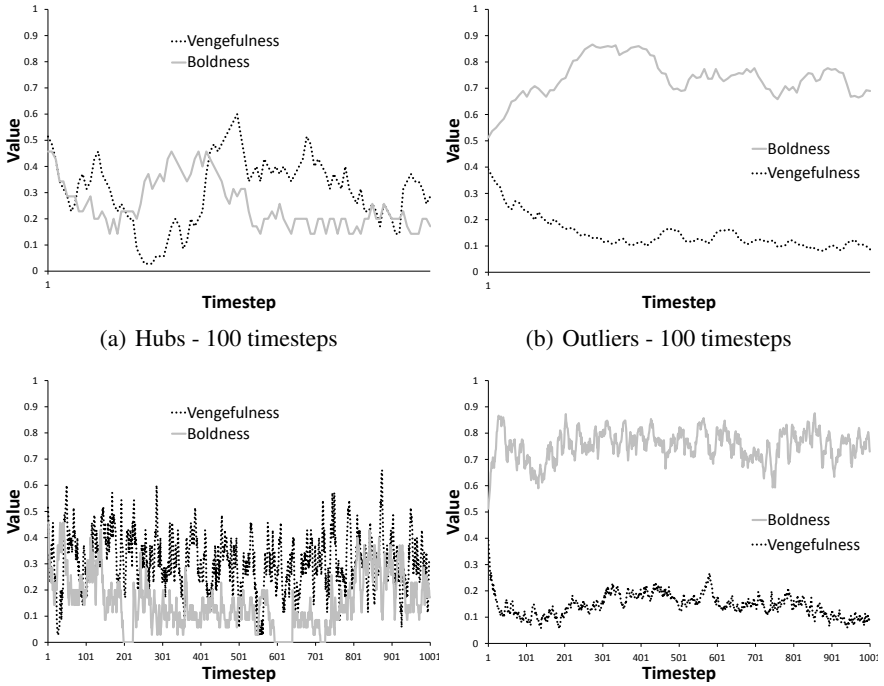


Fig. 4. Universal learning- Overall Result - 1000 Runs, 1000 Agents, 1000 Timesteps

Experiments with this new approach give the results shown in Figure 3.1. Surprisingly, the results indicate norm collapse, as all runs end with high boldness and low vengefulness. By analysing the performance of the different types of agents, we are able to explain this behaviour; we illustrate by reference to a sample run for a hub in Figures 5(a) and 5(c), and a sample run for an outlier agent shown in Figures 5(b) and 5(d).

Outliers have few connections, but are connected to one or more hubs. When agents punish others, they pay an enforcement cost but risk metapunishment when they do not. However, since these outliers have very low connectivity, the risk of metapunishment is also very low, so they avoid punishing others and vengefulness consequently decreases. Metanorms are thus not effective here because of the lack of connectivity between agents. Outliers thus always have high boldness and low vengefulness levels. In addition, as we will see, the vengefulness of hubs also drops and is never higher than average, so agents can defect and gain benefit, without being punished by hubs. Outliers thus increase their boldness, causing norm collapse in the whole population.

In contrast to outliers, hubs are highly connected and apply punishments to many others, incurring high enforcement costs. To address this, they decrease their vengefulness, resulting in metapunishment from the many nodes to which they are connected, in turn causing hubs to increase vengefulness (but only to a mid-range level). In addition, because of the high boldness of outliers, there is a high rate of defection in the population, causing oscillation between mid-range and low vengefulness for the duration of the run. Boldness of hubs is kept low, however, due to the amount of punishment that the hubs are exposed to. Values for vengefulness and boldness are shown Figure 5(c).



(a) Hubs - 100 timesteps (b) Outliers - 100 timesteps

(c) Hubs - 1000 timesteps (d) Outliers - 1000 timesteps

Fig. 5. Universal learning - Sample Run

3.2 Connection-Based Observation

Axelrod's original model considers a probability of being seen, and in the context of a fully connected network, this may be a reasonable basis on which to base a model. However, in the kinds of topologies we are concerned with, such as those that reflect the situations in peer-to-peer (P2P) networks or wireless sensor networks, for example, observation of the behaviour of others arises from the direct connection between agents. Thus, if a peer A is connected to another peer B, then A may be able to observe all communication from B. As a result, if B defects by, for example, not sharing files in the case of a file-sharing P2P network, this can be observed by A. To reflect this property in our model, Axelrod's probability of being seen requires replacing with the notion that each agent observes all actions of its direct neighbours. This modification to the model gives rise to rather different results. In particular, the results of running the model on a scale-free network, in Figure 3.2, show that all runs end in low boldness and low vengefulness, indicating that defection is very rare in the population because of the low boldness. In addition, punishment is not common since agents rarely punish defectors, due to their low vengefulness. To understand this better, the results of a 1000 timestep run, for outliers and hubs, are shown in Figures 7(d) and 7(c), respectively.

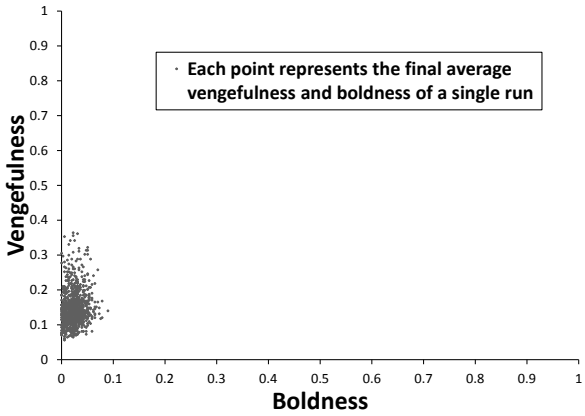


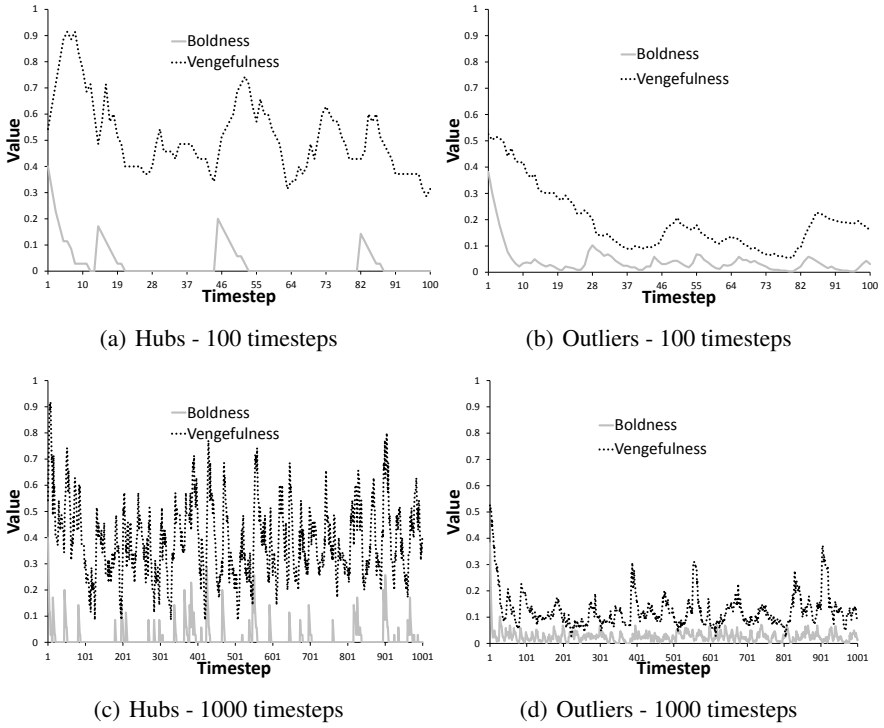
Fig. 6. Connection-Based Observation - Overall Result - 1000 runs, 1000 agents, 1000 timesteps

More specifically, Figures 7(b) and 7(d) show that outliers start the run by decreasing both vengefulness and boldness to a low level where they remain, with some small degree of fluctuation. Figures 7(a) and 7(c) suggest that hubs start the run by increasing their vengefulness to a high level and decreasing their boldness to a very low level. After a few timesteps, vengefulness decreases to a mid-range level, from which it decreases further to a low level. However, it does not stabilise there, since it moves up again, and this pattern is repeated throughout the run. Similarly, boldness initially decreases to zero and then jumps to a low level, before decreasing back to zero. Hubs thus have a fluctuating mid-range level of vengefulness, and a very low level of boldness.

There are two distinctive features that can be observed here, in contrast to the results obtained by the universal learning approach. First, hubs reach a high level of vengefulness, which is limited to mid-range vengefulness in the previous approach. This is mainly because the new technique raises the action observation probability to 100%, which allows a high possibility for metapunishment to occur and, as a result, forces hubs to increase their vengefulness to a high level. However, as before, this does not persist because of the high enforcement cost observed with such a high level of vengefulness. Second, the boldness of outliers is low here, mainly due to the combination of the high vengefulness among hubs and the 100% defection observation, which together produce sufficient punishments to force outliers to decrease their boldness.

4 Dynamic Policy Adaptation

As we have seen, universal learning has a negative impact on results, causing boldness to increase and vengefulness to decrease. However, a more important weakness is that the learning rate is uniform in the face of differing punishment levels: all agents use the same learning rate, regardless of how much utility gain or loss they suffer. Thus, for example, an agent that incurs a punishment score of -10 must modify its vengefulness to exactly the same degree as another agent whose punishment score is -999. While



(a) Hubs - 100 timesteps (b) Outliers - 100 timesteps

(c) Hubs - 1000 timesteps (d) Outliers - 1000 timesteps

Fig. 7. Connection-Based Observation - Sample Run

the direction of change is appropriate, the degree of change does not reflect the severity of the sanction; a more appropriate approach would change policy in line with performance. In this view, a very badly performing agent should modify its policy much more significantly than one that performs better. Dynamic policy adaptation can address this, bringing about changes to vengefulness and boldness that reflect performance. The key idea here is to measure the *level* of performance rather than just the *direction*, through comparison of an agent's actual utility, or *score* in our terms, and the maximum or minimum that could be obtained. We apply this to boldness and vengefulness in turn, but first introduce some notation. Let NDD be the number of available defection decisions, where each agent has multiple chances to defect in a single round (as specified earlier), NB_i be the number of i's neighbours, T be the utility gained from a single defection, and PC be the punishment cost representing the utility lost from being punished.

4.1 Boldness

In terms of boldness, the relevant part of the total score is the *defection score*, which can be either positive or negative, requiring consideration of both maximum and minimum possible values. The maximum possible defection score $MaxDS_i$ arises when an agent i always defects but is never punished, and the minimum defection score arises when

the agent always defects and is always punished by all of its neighbours, as follows.

$$MaxDS_i = NDD \times T \tag{1}$$

$$MinDS_i = NDD \times (T + (NB_i \times PC)) \tag{2}$$

Then, in order to determine the degree of change to an agent's boldness, we must consider three different situations. First, when the defection score is positive (so that boldness should increase), the degree of change is determined by dividing the obtained defection score by the maximum possible defection score. Second, when it is negative, (so that boldness should decrease), the obtained defection score is divided by the minimum possible defection score. Finally, if the defection score is zero, no change is required.

$$FactorB_i = \begin{cases} \frac{DS_i}{MaxdDS_i} & \text{if } DS_i > 0 \\ \frac{DS_i}{MinDS_i} & \text{if } DS_i < 0 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Given this, we now need to determine how $FactorB$ can be used to change an agent's policy. In order to avoid dramatic policy movements that could lead to violent fluctuations, we limit the change that can be applied to a maximum value. In this case, the maximum is the difference between two levels as in Axelrod's original model, of $\frac{1}{7}$. Thus, an agent modifies its boldness in line with its DS, as follows, so that it can maximally change its boldness by one level (or by $\frac{1}{7}$) when $FactorB$ is 1.

$$B_i = B_i + \begin{cases} \frac{1}{7} \times FactorB_i & \text{if } DS_i > 0 \\ -\frac{1}{7} \times FactorB_i & \text{if } DS_i < 0 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

4.2 Vengefulness

An agent modifies its vengefulness depending on whether it is valuable to punish others, determined by comparing the utility lost from punishing others (the punishment score, PS) against and the utility lost from not punishing them (the punishment omission score POS). If PS is greater than POS, agents increase vengefulness and decrease it otherwise. Clearly, the magnitude of this difference between these two values gives an indication of the degree of change that should be applied to vengefulness. For example, if PS is -24 and POS is -20, then the degree of decrease to V should be significantly lower than when PS is -600 and POS is -20. We call this difference $DiffV$:

$$DiffV_i = |PS_i - POS_i| \tag{5}$$

Since $DiffV$ is 1 or more (when the values are not equal), it cannot be used directly to update an agent's V value, because V must always lie between 0 and 1. It must thus be *normalised* so that it can be applied to V, for which we use a scaled value, $FactorV$; this is determined by dividing $DiffV$ by the minimum of PS and POS. Since both PS and POS are negative, their absolute value is used to obtain a positive value:

$$FactorV_i = \frac{DiffV}{|\min\{PS_i, POS_i\}|} \tag{6}$$

While this always produces a value between 0 and 1, it does not provide the same value for the same magnitude of difference. For example, if PS is -14 and POS is -20, we want $FactorV_i$ to be the same as when PS is -6 and POS is 0. We can achieve this by replacing the maximum of PS and POS with the maximum possible difference between PS and POS. This maximum difference is the difference from 0 (when there is no cost at all from punishing or from not punishing) to the greatest possible magnitude of PS or POS. In what follows, HPS represents the highest punishment score (the maximum in magnitude, and lowest in numerical terms — we use HPS to indicate the *highest* score to avoid ambiguity of minimum and maximum) that can be received by an agent punishing all of its neighbours for defection, and metapunishing all of its neighbours for not punishing all of their neighbours for defection.

To determine the value of HPS we need to consider both the punishment enforcement cost and the metapunishment enforcement cost. First, the highest (maximum in magnitude, but minimum numerically) *punishment* enforcement cost ($HPEC$) arises when all of an agent's neighbours defect and the agent punishes all of them:

$$HPEC_i = NDD \times NB_i \times EC \tag{7}$$

where EC is the enforcement cost of a single punishment. Similarly, the highest *metapunishment* enforcement cost ($HMPEC$) arises when all of an agent's neighbours do not punish all of their neighbours for defecting, and the agent metapunishes all of them:

$$HMPEC_i = NDD \times NBB_i \times EC \tag{8}$$

where NBB_i is the total number of neighbours of all of agent i's neighbours. HPS is thus defined as the sum of these two scores:

$$HPS_i = HPEC_i + HMPEC_i \tag{9}$$

In the same way, $HPOS$ is the highest (greatest in magnitude, lowest numerically) score that can be obtained when an agent does not punish any defectors, but is metapunished by all of its neighbours.

$$HPOS_i = NDD \times NB_i \times (NB_i - 1) \times PC \tag{10}$$

where the maximum number of defectors is all of an agent's neighbours (NB), the maximum number of metapunishers is the same but excluding the defecting agent, and PC is the punishment cost obtained from being metapunished (which is the same as for simply being punished). Given this, $FactorV$ can be calculated by dividing $DiffV$ by one of these values, as follows. If punishing brings a greater utility reduction than not punishing ($PS < POS$), then we use the highest punishment score HPS. Conversely, if $PS > POS$, then we use the highest punishment omission score $HPOS$. If there is no difference, then there is no change and $FactorV$ is equal to 0.

$$factorV_i = \begin{cases} \frac{DiffV_i}{HPS_i} & \text{if } POS_i > PS_i \\ \frac{DiffV_i}{HPOS_i} & \text{if } POS_i < PS_i \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

This guarantees that the change made to V is always the same given the same difference in scores, since both HPS and $HPOS$ are fixed for each agent. Moreover, this approach allows hubs to change much less quickly than outliers, because the highest (maximum in magnitude) scores for hubs are much higher than for outliers, so that the results achieved by using $FactorV$, and dividing by the difference in scores obtained for hubs, is much less than for outliers. As the learning algorithm suggests, an agent increases vengefulness when it finds that not punishing is worse than punishing, and it decreases vengefulness when the converse is true.

$$V_i = V_i + \begin{cases} \frac{1}{7} \times FactorV_i & \text{if } |PS_i| < |POS_i| \\ -\frac{1}{7} \times FactorV_i & \text{if } |PS_i| > |POS_i| \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

4.3 Example

To illustrate, assume that a hub A is connected to 20 other agents, and that an outlier B is connected to only 2 other agents (one being a hub). Like Axelrod's seminal experiments and without loss of generality, let $NDD = 4$ for all agents, since every agent has 4 chances to defect in each round. $EC = -2$ and is the same for all agents. Similarly, $PC = -9$ and again is the same for all agents. The temptation value for all agents, received when they defect, is $T = 3$. Finally, suppose that A's neighbours have 50 other distinct neighbours in total (summed over all neighbours), while B's neighbours have 20 other distinct neighbours (again, over all). This is summarised in Table 1. Given these values, we can determine the relevant values needed as follows. Starting with defection scores and from Equations 1 and 2 respectively, we obtain the following:

$$\begin{aligned} MaxDS_A = MaxDS_B &= 4 \times 3 = 12 \\ MinDS_A &= 4 \times (3 + (20 \times -9)) = -708 \\ MinDS_B &= 4 \times (3 + (2 \times -9)) = -60 \end{aligned}$$

In terms of punishment values, from Equations 7, 8 and 9, we obtain the following:

$$\begin{aligned} HPEC_A &= 4 \times 20 \times -2 = -160 \\ HMPEC_A &= 4 \times 50 \times -2 = -400 \\ HPS_A &= -160 - 400 = -560 \\ HPEC_B &= 4 \times 2 \times -2 = -16 \\ HMPEC_B &= 4 \times 20 \times -2 = -160 \\ HPS_B &= -16 - 160 = -176 \end{aligned}$$

Punishment omission scores using Equation 10 are as follows:

$$\begin{aligned} HPOS_A &= 4 \times 20 \times 19 \times -9 = -13680 \\ HPOS_B &= 4 \times 2 \times 1 \times -9 = -72 \end{aligned}$$

Using this information (Table 1), we can determine the decisions for specific situations. For example, at the start of each run, the population has average mid-range

Table 1. Example values for Agents A and B

Agent	Pos	NB	NBB	MinDS	MaxDS	LevB	HPS	HPOS	LevV
A	Hub	20	50	-708	12	1/7	-560	-13680	1/7
B	outlier	2	20	-60	12	1/7	-176	-72	1/7

boldness and vengefulness (because of the uniform distribution function to generate initial policies). Now, suppose that both A and B also have mid-range boldness and vengefulness. If, after one round, both A and B defected twice (out of their four opportunities to defect), they each gain twice the temptation value T. However, since A is a hub, suppose it is punished 22 times, much more than B, which is only punished twice. This is because the defection score of a hub with mid-range boldness is typically much worse than that of a similar outlier, mainly due to the difference in their number of neighbours, and the midrange vengefulness in the population. Thus, A has a defection score of 2×3 from defecting, plus $22 \times -9 = -198$ from being punished, giving $DS_A = -192$. Similarly, $DS_B = ((2 \times 3) + (2 \times -9)) = -12$.

Given these defection scores, the degree of change that each agent applies to its boldness can be calculated as follows. First, from Equation 3, $FactorB_A = \frac{-192}{-708} = 0.3$ and $FactorB_B = \frac{-12}{-60} = 0.2$. Now, using Equation 4, and since both DS_A and DS_B are negative, B_A is decreased by $0.3 \times \frac{1}{7} = 0.04$, and B_B by $0.2 \times \frac{1}{7} = 0.03$.

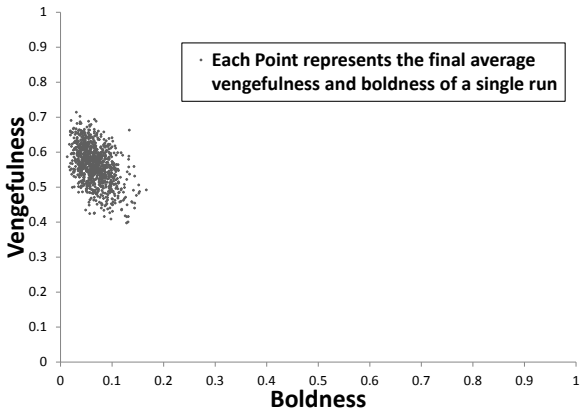


Fig. 8. Dynamic Policy Adaptation - Overall Result - 1000 Runs, 1000 Agents, 1000 Timesteps

In addition, if A punishes 35 other agents and metapunishes 16 more, and B punishes 10 other agents and metapunishes 4 more, their punishment scores are determined by multiplying the number of punishments issued by the enforcement cost EC: $PS_A = ((35{+}16){\times}{-}2) = -102)$ and $PS_B = ((10{+}4){\times}{-}2) = -28)$. Then, if A has spared 27 defectors and has been metapunished 6 times for each instance of omitting punishment, and B has spared only one defector and been metapunished just once, the

punishment omission scores are calculated by multiplying the number of metapunishments by the punishment cost PC, as follows: $POS_A = (27 \times 6 \times -9) = -1458$ and $POS_B = (1 \times 1 \times -9) = -9$. Thus, by Equation 11, $FactorV_A = \frac{|-102-(-1458)|}{13680} = 0.1$ and $FactorV_B = \frac{|-28-(-9)|}{96} = 0.2$. Then, since $PS_A > POS_A$, A increases its vengefulness V_A by $0.1 \times \frac{1}{7} = 0.014$ according to Equation 12). Similarly, since $PS_B < POS_B$, B decreases its vengefulness by $0.2 \times \frac{1}{7} = 0.03$.

4.4 Experimental Results

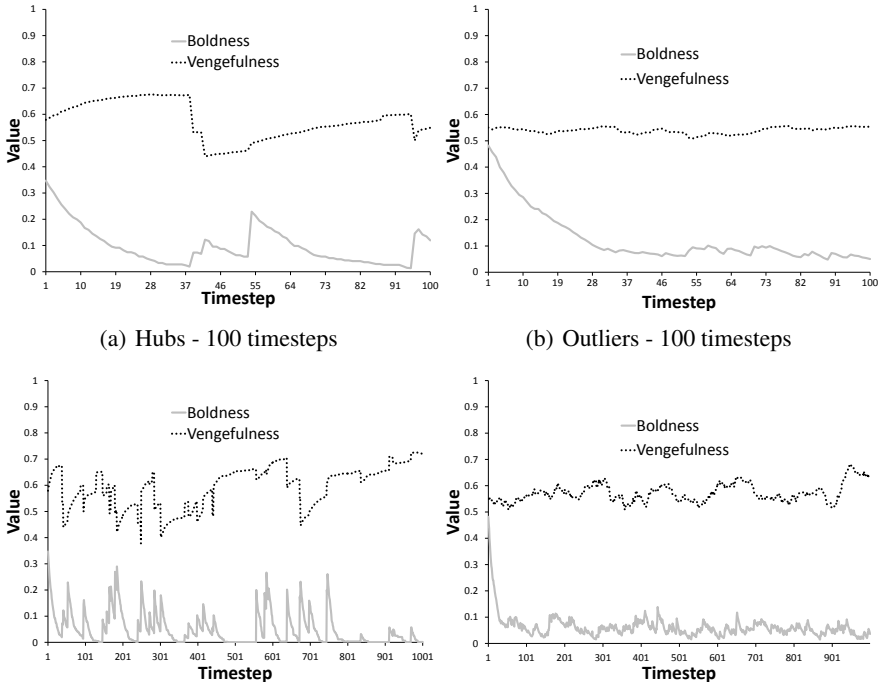
To determine the effect of introducing dynamic policy adaptation, we ran experiments, similar to the previous experiments, on the new model, and giving the results shown in Figure 4.3. As can be seen in the figure, all runs result in populations with low average boldness and moderate vengefulness. As before, more detail on the evolution of average boldness and vengefulness for hubs and outliers was provided by examining runs of individual agents, as shown in Figure 9 , which confirm that outliers converge to a state of low boldness and moderate vengefulness consistently, while hubs do so with intermittent deviations. As before, hubs increase vengefulness and decrease boldness, though much more slowly now. However, at regular intervals, there are sudden increases to boldness, accompanied by a change in vengefulness, as a result of the exploration of the algorithm. This phenomenon occurs in all models in this paper, and is visible here due to the limited number of timesteps, but has no impact on the results of the dynamic policy adaptation.

5 Conclusion

Norm emergence is an important and valuable phenomenon that has applications to self-organising systems such as peer-to-peer networks or wireless sensor networks in which there is no interference from any central or outside authority. While there has been much work on this phenomenon (as discussed earlier), and even some on its application to different topological structures, there has been inadequate consideration of how to establish norms in scale-free networks. Indeed, some mechanisms have been shown not to succeed in these topologies. In response, this paper provides an effective means of overcoming the problems arising from asymmetric connections of hubs and outliers.

In particular, our results show that in scale-free networks, Axelrod's basic metanorm model is not effective, nor is Mahmoud et al.'s attempt to overcome this for other topologies. Our simulations suggest that poorly connected agents receive little discouragement from defecting while hubs are discouraged from enforcing norms through high enforcement costs. In response, we have modified the experimental setting to be more consistent with the nature of distributed systems of partially connected nodes, bringing an even more serious breakdown in norm emergence, but also subsequently addressed this through a dynamic policy adaptation mechanism. In this way, agents are able to change their policy in proportion to the punishments they receive, allowing them to adapt proportionally, and to maintain the policy values that sustain norm establishment.

In terms of future work, we plan to conduct a more detailed analysis of the effect of different levels of the probability of observation on the results of the model. In addition,



(a) Hubs - 100 timesteps (b) Outliers - 100 timesteps

(c) Hubs - 1000 timesteps (d) Outliers - 1000 timesteps

Fig. 9. Dynamic Policy Adaptation - Sample Run

we aim to develop an efficient adaptive punishment approach that allows punishment to be applied according to the specific case at hand, so that agents with different degrees of defection will be punished accordingly. This is important so that the punishment is no greater than needed, and will not overly constrain behaviour (limiting the functionality of the underlying system, which is undesirable). We believe that the dynamic policy adaptation technique that has been introduced in this paper provides a solid grounding for just such an adaptive punishment approach.

References

1. E. Adar and B. A. Huberman. Free riding on gnutella. *First Monday*, 5(10), 2000.
2. R. Axelrod. An evolutionary approach to norms. *The American Political Science Review*, 80(4):1095–1111, 1986.
3. A. L. Barabasi and R. Albert. Emergence of Scaling in Random Networks. *Science*, 286(5439):509–512, 1999.
4. A. P. de Pinninck, C. Sierra, and W. M. Schorlemmer. Friends no more: norm enforcement in multiagent systems. In *Proceedings of the Sixth International Joint Conference on Autonomous Agents and Multi-Agent Systems*, pages 640–642, 2007.
5. J. Delgado, J. M. Pujol, and R. Sangüesa. Emergence of coordination in scale-free networks. *Web Intellgence. and Agent Systems*, 1:131–138, 2003.

6. J. M. Epstein. Learning to be thoughtless: Social norms and individual computation. *Computational Economics*, 18(1):9–24, 2001.
7. F. Flentge, D. Polani, and T. Uthmann. Modelling the emergence of possession norms using memes. *Journal of Artificial Societies and Social Simulation*, 4(4), 2001.
8. J. M. Galán, M. M. Latek, and S. M. M. Rizi. Axelrod's metanorm games on networks. *PLoS ONE*, 6(5):e20474, 2011.
9. S. Mahmoud, J. Keppens, M. Luck, and N. Griffiths. Norm establishment via metanorms in network topologies. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, volume 3, pages 25–28, 2011.
10. S. Mahmoud, J. Keppens, M. Luck, and N. Griffiths. Overcoming omniscience in axelrod's model. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, volume 3, pages 29–32, 2011.
11. B. T. R. Savarimuthu, S. Cranefield, M. Purvis, and M. Purvis. Norm emergence in agent societies formed by dynamically changing networks. In *Proc. 2007 IEEE/WIC/ACM International Conference on Intelligent Agent Technology*, pages 464–470, 2007.
12. B. T. R. Savarimuthu, M. Purvis, M. Purvis, and S. Cranefield. Social norm emergence in virtual agent societies. In *Declarative Agent Languages and Technologies VI*, volume 5397 of *Lecture Notes in Computer Science*, pages 18–28. Springer, 2009.
13. O. Sen and S. Sen. Effects of social network topology and options on norm emergence. In *Coordination, Organizations, Institutions and Norms in Agent Systems V*, volume 6069 of *Lecture Notes in Computer Science*, pages 211–222. 2010.
14. Y. Shoham and M. Tennenholtz. On social laws for artificial agent societies: off-line design. *Artificial Intelligence*, 73(1-2):231–252, 1995.
15. D. Villatoro, S. Sen, and J. Sabater-Mir. Topology and memory effect on convention emergence. In *Proceedings of the 2009 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technologies*, pages 233–240.
16. A. Walker and M. Wooldridge. Understanding the Emergence of Conventions in Multi-Agent Systems. In V. Lesser, editor, *Proceedings of the First International Conference on Multi-Agent Systems*, pages 384–389. MIT Press, 1995.