nature communications

Article

https://doi.org/10.1038/s41467-022-34102-1

An AI approach for managing financial systemic risk via bank bailouts by taxpayers

Received: 11 April 2022

Accepted: 13 October 2022

Published online: 17 November 2022

Check for updates

Daniele Petrone[1], Neofytos Rodosthenous[2] & Vito Latora[1,3,4]

Bank bailouts are controversial governmental decisions, putting taxpayers' money at risk to avoid a domino effect through the network of claims between financial institutions. Yet very few studies address quantitatively the convenience of government investments in failing banks from the taxpayers' standpoint. We propose a dynamic financial network framework incorporating bailout decisions as a Markov Decision Process and an artificial intelligence technique that learns the optimal bailout actions to minimise the expected taxpayers' losses. Considering the European global systemically important institutions, we find that bailout decisions become optimal only if the taxpayers' stakes exceed some critical level, endogenously determined by all financial network's characteristics. The convenience to intervene increases with the network's distress, taxpayers' stakes, bank bilateral credit exposures and crisis duration. Moreover, the government should optimally keep bailing-out banks that received previous investments, creating moral hazard for rescued banks that could increase their risk-taking, reckoning on government intervention.

In times of crisis, as during the recession of 2008 or the economic disruption triggered by the COVID-19 pandemic, governments face difficult decisions regarding bailing-out strategically important companies. In particular, large banks and other financial institutions are critical for the stability of the financial system and are closely monitored by central banks and government departments. It is nowadays widely understood, that the stability of the financial system cannot be assessed focusing exclusively on each individual financial institution. The interconnections and interactions between financial institutions are at least as important in contributing to the overall dynamics (see refs. 1–5). It thus requires a broader approach to manage the risk that a considerable part of the financial system is disrupted (systemic risk). A number of regulatory boards and committees, such as the US Financial Stability Oversight Council, the European Systemic Risk Board and the Bank of England's Financial Policy Committee, have been created in order to identify, monitor and take actions that can remove or reduce the systemic risk. They are also tasked to look for new methodologies and ideas from different disciplines to deepen their understanding of the complex phenomena involved in financial crises.

For example, in order to rescue the Royal Bank of Scotland (RBS) in 2008–2009, the UK government became the majority shareholder of the bank, purchasing shares for a total of £45.5 billion, according to ref. 6. The government achieved its objective to stabilise the financial system and no depositor in UK banks lost any money. However, the cost for taxpayers has been estimated by the Office for Budget Responsibility[7] to be in the region of £27 billion. The price of RBS shares plummeted after the purchase and the government has since sold part of its investment at a loss. Was this governmental intervention value for money? The National Audit Office (NAO)[8], UK's public spending watchdog, released a report on maintaining financial stability across the UK's banking system, analysing the governmental support to the banking sector and concluded that: If the support measures had not been put in place, the scale of the economic and social costs, if one or more major UK banks had collapsed, is difficult to envision; the

[1]School of Mathematical Sciences, Queen Mary University of London, E1 4NS London, UK. [2]Department of Mathematics, University College London, WC1E 6BT London, UK. [3]Dipartimento di Fisica ed Astronomia, Università di Catania and INFN, Catania I-95123, Italy. [4]Complexity Science Hub Vienna, A-1080 Vienna, Austria. e-mail: n.rodosthenous@ucl.ac.uk

support provided to the banks was therefore justified, but the final cost to the taxpayer of the support will not be known for a number of years.

A more recent example is the COVID-19 pandemic, which has had a devastating effect on economies worldwide with further forthcoming effects not fully observed yet. The quarter-over-quarter change in the US GDP fell by 31.2% in the second quarter of 2020, while the Office for National Statistics estimated that the UK economic output fell by 9.9% during the year 2020, the largest annual fall on record. The banks have so far weathered the storm, aided by improved regulations and macroprudential measures introduced in the aftermath of the global financial crisis of 2008–2009. However, no one can predict if the financial system can withstand a series of bankruptcies in the property, aviation, creative, tourism and hospitality sectors, that might ensue as the accommodating monetary policies of central banks are tapered due to the dramatic surge of inflation.

As the main concern is the systemic risk that a default entails, a network model is essential (see, e.g. refs. 9–12 for recent reviews of the financial systemic risk literature). The nodes of the network are banks or other financial institutions and their links represent mutual exposures. The connections between financial institutions can then transfer the distress amongst them (see, e.g. refs. 1, 13, 14). There is a very large literature borrowing techniques from network science (see, e.g. refs. 15, 16) and successfully applying them to the study of network resilience to external shocks in order to provide useful analyses of financial systemic risk (see, e.g. refs. 13, 17, 18). There is also a vast literature on governmental interventions in financial institutions, which spans across many different directions, such as post-bailout bank performances[19], the bailout effects on the underwriting business[20], on market discipline[21] and on sovereign risk[22], as well as the interplay amongst bailouts, banks' risk profile and national regulation[23–26]. Our work though is related to a branch of the financial systemic risk literature that analyses interventions to limit the effects of financial crises. In particular, the relevance of bailout actions in mitigating the contagion during the financial crisis of 2008 is evidenced in ref. 27, while various network models of government interventions are proposed in refs. 28–32.

Although bank bailouts are among the most critical decisions a government can take, very few studies have addressed quantitatively the problem of assessing their convenience for the taxpayers, as we do in this paper. To be more precise, the main differences with existing literature are that: (a) our framework does not require starting the analysis with a set of banks already in default or about to default without a government intervention, hence we allow preventive actions before the network is compromised; (b) the previous literature considers the optimization of (multiple variations of) functions based on social costs, system wealth and taxpayers' bailout money—we instead focus on minimising the loss for taxpayers during the crisis, irrespective of the size of the system's overall wealth, which may not be directly linked to the taxpayers' interests; (c) our modelling approach and framework fills the gaps in literature, by allowing the possibility that our dynamic network can be controlled by governments, at each and every time step, via injections of additional capital in financial institutions.

In particular, we propose a mathematical framework that allows for a quantitative comparison between different potential investments in financial institutions by the government. Our framework is based on the three following building blocks: (a) a dynamical network model of the financial system that describes the contagion mechanism between financial institutions (modelled as an increase in the probability of default of banks that have claims on failed institutions); (b) a set of allowed government interventions to control the network (investments in the capital of distressed banks); and (c) a quantitative way to assess government actions or inaction at each time step (using artificial intelligence techniques). Our main aim is to address the eventuality that a government needs to decide whether to bailout a financial institution or let it fail as its insolvency becomes more and more likely.

The contagion mechanism that we use is the impact that a bank default has on other banks. The impact can be due to: (a) direct losses from cross ownership and bilateral credit exposures (for example loans, see, e.g. refs. 33, 34), or (b) indirect losses due to fire selling of assets by defaulting banks, that would lower the market value of similar assets in the balance sheet of non-defaulting financial institutions (see, e.g. ref. 35). In all, the impact of defaults would lower the capital buffer of affected banks, thus weakening the entire network and its ability to withstand future shocks. In our model, this is accounted for by an increase of the probability of default (PD) per unit time of the banks that have claims towards the defaulted institutions. Such an existence of a PD as a characteristic of the nodes of the bilateral credit exposures network has only recently been introduced in systemic risk evaluation studies (see, e.g. refs. 36, 37).

One of our main novelties with respect to the aforementioned models is that we further allow for the network to be controlled by a government via investments in the capital of banks. Such an investment would, conversely to defaults, decrease the banks' PD upon receiving the additional capital. The nodes' PD thus eventually allows us to follow the evolution in time of our (controlled) dynamical model as there is a well-defined length of time during which the government can intervene to control the network. We provide the connection between the changes in each node's PD and the changes in the amount of capital due to the impact from defaults or governmental investments via the credit risk model introduced by Merton[38]. Then, given each node's PD, as well as the likelihood of more than one nodes defaulting simultaneously (during the same time step), which we describe by a Gaussian latent variable model, we follow the stochastic evolution of the network in time via a multi-period Monte Carlo simulation.

Even though government investments decrease the banks' PD, they also increase the potential loss of this additional capital for the government (and taxpayers) in case of default. This creates a trade-off for the decision-makers. The main aim of the government is therefore to answer the questions of whether to invest in financial institutions at each time step, which financial institution(s) to invest in and how much to invest, in order to achieve the minimum expected taxpayers' loss during a crisis.

To that end, we model the system's evolution as a Markov Decision Process (MDP) (see ref. 39), where the actions (controls) are government investments in the capital of the banks at each time step, and the dynamics and negative rewards (losses) are linked to the financial network dynamics, each node's (controlled) probability of default, total asset and previous government investments (the sequence of governmental stochastic controls). However, our MDP is both challenging to define in this setup and (even more) challenging to solve, given its following main characteristics: (a) the MDP state definition is remarkably complex since it depends on all the parameters of the network at each specific time; (b) the low probability of default of each node in the network translates in a high probability of not receiving any reward signals to learn the best action, (c) the enormous number of successor states (even in very simple networks) would normally make standard computations impossible and standard methods non-feasible.

In order to overcome these challenges, we develop an artificial intelligence technique that uses a variation of the Fitted Value Iteration algorithm (see, e.g. refs. 40, 41) with bespoke characteristics that are uniquely constructed to solve our MDP. To be more precise, we: (a) devise a particular value function parametrisation, representing the sum of the expected direct losses per node and remaining time steps; (b) implement a learning process backwards in time from the end of the stochastic episode (financial crisis) where the value function is known to be zero; and (c) use an ingenious duality between the dynamics of the financial network (our nodes' default modelling) and the MDP rewards and transition probabilities, to reduce drastically the

number of terms in many critical expressions—in particular, we devise a technique for rewriting these expressions in terms of (the remarkably smaller) number of non-defaulted nodes rather than successor states (according to standard theory), hence allowing their computation. Our methodology allows us to assess the optimality of government decisions—no investment versus different types and amounts of investment—and conclude the optimal government actions per time step and state of the network.

The introduced framework has a high potential for becoming an important tool for central banks and governments, whose budget, data and resources could allow for a professional calibration of our model. The mathematical assessment and dynamic optimisation of bank bailout decisions from a taxpayer's standpoint could be a valuable quantitative tool in their diverse toolbox (make new bailout decisions, apply on selected bailout cases to evaluate results, learn from past experiences and actual happenings). Furthermore, our proposed methodology could have not only a practical impact to assess government interventions, but also a significant impact on the scientific community aiming at tackling problems of stochastic control in dynamic networks with few reward signals, as the one we solve in this paper.

Results

One of the main results of our paper is the introduction of a mathematical framework, that allows governments to assess the convenience to intervene with bailout investments in distressed banks' equity and optimise their decisions. In the following subsections, we first present the network model of financial institutions, its dynamics and contagion mechanism, and then propose an MDP based on the network, which will be used to model government interventions on bailed-out financial institutions. After formulating the problem, we proceed with a methodology for solving the MDP and some of the mathematical technicalities. Finally, we implement our framework and methodology on two case studies and present our findings.

Network of financial institutions

We consider a network whose set $\mathcal{I} = \{1, \ldots, N\}$ of nodes represents financial institutions. Each node $i \in \mathcal{I}$ is characterised at time t by a probability of default $PD_i(t) \in (0,1)$ per time interval Δt, a total asset $W_i(t)$ and an equity $E_i(t)$, that is the capital used by node i as a buffer to withstand financial losses, satisfying $E_i(t) \le W_i(t)$.

The edge (i, j) of the network represents the exposure of node i to the default of node j where $i \neq j \in \mathcal{I}$. Each edge (i, j) is associated to a numerical value w_{ij} which depends on the contagion channels considered. For example, we can consider only credit exposures or also the impact due to fire sales of common assets. Regarding credit exposures, most of the times only aggregated values are available, e.g. the total amount of inter-banking assets and liabilities for each node, and in such cases, bespoke algorithms are used to infer the network of bilateral exposures (see, e.g. refs. 37, 42). To take into account government interventions aimed at limiting the overall losses, we use an adaptation of the PD model introduced in ref. 37 by extending it to allow the possibility for the nodes to incur also positive shocks, via investments in the nodes, rather than just negative shocks due to the default of other nodes. The focus of this paper is also radically different from the one in ref. 37, which focuses on the losses sustained by private investors, since we are here exclusively interested in the losses incurred by the taxpayers. In the following, we will measure the time in discrete time steps that are multiples of Δt, i.e. $t+1$ is equivalent to $t + \Delta t$.

We define the total impact $I_i(t)$ on node i at time t, due to the default of other nodes $j \in \mathcal{I} \setminus \{i\}$ in the network and their exposure w_{ij}, by

$$I_i(t) := \sum_{j \in \mathcal{I} \setminus \{i\}} w_{ij} \delta_j(t), \quad \text{for all } i \in \mathcal{I}, \tag{1}$$

where

$$\delta_j(t) = \begin{cases} 1, & \text{if node } j \text{ defaults at time } t, \\ 0, & \text{otherwise.} \end{cases}$$

The mechanism by which defaults will occur at each time step t, yielding $\delta_\cdot(t) = 1$, will be constructed towards the end of this section using all network information, including the probabilistic framework up to time t. The impact $I_i(t)$ represents a loss for the total asset W_i, which in turn decreases also the equity E_i of node i, hence reducing their value at time $t+1$. This can be seen from the accounting equation for each node i, namely

$$W_i(t) = E_i(t) + B_i(t), \tag{2}$$

which states that the total asset W_i is always equal, at all times, to the equity E_i plus the total liability B_i. Note that B_i is not affected by the losses as it is comprised of loans from other banks, deposits, etc., that are due in full unless the bank i defaults. Hence, we have

$$\Delta W_i(t) = \Delta E_i(t), \tag{3}$$

where we define $\Delta X_i(t) := X_i(t+1) - X_i(t)$. Taking into account also the potential increase $\Delta J_i(t)$ in the cumulative investment $J_i(t)$ of the government in node i up to time t, which will in turn increase the values of the total asset W_i and equity E_i of node i at time $t+1$, we can write

$$W_i(t+1) = W_i(t) - I_i(t) + \Delta J_i(t) \quad \text{and} \quad E_i(t+1) = E_i(t) - I_i(t) + \Delta J_i(t). \tag{4}$$

The probability of default $PD_i(t)$ of node i is increased by the impact $I_i(t)$ at time t, since part of the capital buffer (equity E_i) is lost, and decreased by the potential investment $\Delta J_i(t)$, which in turn grows the capital buffer. In order to model the effect of the impact $I_i(t)$ and potential investment $\Delta J_i(t)$ on $PD_i(t)$, we use here the credit risk model introduced by Merton[38]. Alternatively, it is possible to use the first passage model introduced by Black and Cox[43]. The implied probability of default PDM is therefore calculated as a function of the parameters of each node:

$$PDM(W, E, \mu, \sigma) := 1 - \Phi\left(\left(\log\frac{W}{W-E} + \mu - \frac{\sigma^2}{2}\right)\Big/\sigma\right), \tag{5}$$

where the term $W-E$ represents the total liability B of each bank, Φ is the univariate standard Gaussian distribution, μ is the drift (expected growth rate) and σ is the volatility of the geometric Brownian motion associated to the total asset W in the Merton model. We then use (5) to obtain the probability of default of node i,

$$PD_i(t) := \max\left\{PDM(W_i(t), E_i(t), \mu_i, \sigma_i), PDM_i^{floor}\right\}, \tag{6}$$

where we introduce the fixed number PDM_i^{floor}, whose purpose is to exclude unreasonably low probabilities of default, essentially acting as a lower bound of the PD_i. A lower bound PDM_i^{floor} is necessary, as no matter how well a bank i is capitalised against losses, it can still default due to extreme events such as natural disasters, political revolutions, sovereign defaults, etc. Without PDM_i^{floor}, the government would underestimate the actual probability of default and would tend to invest more capital than it is convenient. As an example of calibration of this parameter, we can follow the standard assumption that the PD_i of a bank $i \in \mathcal{I}$ is greater or equal to the probability of default of the country where it is based in. In this context, the PDM_i^{floor} would be the probability of the country hosting bank i to default on its debt.

Now, if node i loses an amount of capital $I_i(t)$ greater than its capital buffer (equity $E_i(t)$), at some time t, the total asset $W_i(t)$ becomes less than its liability $B_i(t)$ and it is convenient for the

shareholders to exercise their option to default. In practice, when this occurs, we set $PD_i(t+1)=1$ and node i will default at time $t+1$. Moreover, recall that node i may also default at any time t with probability $PD_i(t)$ due to its own individual characteristics given by (6); see also the default mechanism described at the end of this subsection

Now, when node i defaults, we denote by LGD_i the loss given default of node i, which is a fixed number representing the percentage of the cumulative investments J_i on node i by the government, that cannot be recovered after a default. In case of default of node i, we further assume that in addition to the aforementioned loss of investments, the taxpayers' loss L_i is also comprised of a fixed percentage α_i (for convenience) of the total asset W_i of the node i. That is, the taxpayers' overall loss $L_i(t)$ at time t is given by

$$L_i(t) := \alpha_i W_i(t) + LGD_i J_i(t). \tag{7}$$

To complete our framework, we need to specify the probability of more than one default happening during the same time step, given the PD_i of each node i obtained as in (6). For example, if the nodes were independent, the probability of nodes i and j defaulting at the same time step, denoted by $PD_{[ij]}$, would be the product of the individual probabilities PD_i and PD_j. In this paper, we allow nodes to depend on each other and use a Gaussian latent variable model (see, e.g. ref. 44) to calculate the probabilities of simultaneous defaults of two or more nodes. To be more precise, the probability of a finite subset of nodes $\{i,j,k,\ldots\} \subseteq \mathcal{I}$ of the network defaulting at the same time, is given by

$$PD_{[i,j,k,\ldots]} := \int_D \Phi'_N(\mathbf{u};\Sigma)d\mathbf{u}, \tag{8}$$

where Φ'_N is the standardised multivariate Gaussian density function, with zero mean and a symmetric correlation matrix $\Sigma \in [-1,1]^{N\times N}$, given by

$$\Phi'_N(\mathbf{u};\Sigma) := \frac{\exp\{-\frac{1}{2}\mathbf{u}^T\Sigma^{-1}\mathbf{u}\}}{\sqrt{(2\pi)^n\,|\Sigma|}} \tag{9}$$

and $|\Sigma|$ is the determinant of Σ. We further note that the integration domain D in (8) is the Cartesian product of the intervals $[-\infty,\Phi_1^{-1}(PD_l)]$ for each node l that belongs to the set of defaulting nodes $\{i,j,k,\ldots\}$, and the intervals $[-\infty,\infty]$ for the remaining non-defaulting nodes, where Φ_1 is the univariate standard Gaussian distribution.

We are now ready to present the mechanism according to which nodes can default based on their individual characteristics. To be more precise, at each time step t, we first sample values $(x_1,\ldots,x_N)$ of the random vector $X=(X_1,X_2,\ldots,X_N)$ with the multivariate Gaussian distribution of the underlying Gaussian latent variable model mentioned above. Then, we assume that node i defaults according to the rule:

$$x_i < \Phi_1^{-1}(PD_i(t)) \quad \Longleftrightarrow \quad \delta_i(t)=1. \tag{10}$$

The banks bailout problem as a Markov Decision Process

In this subsection, we describe the government decisions of bailing out banks as a Markov Decision Process (MDP) driven by our framework described in the previous subsection. We firstly assume that the government estimated that the crisis will likely be over at time M, where each time step could be interpreted to reflect the contagion effect, which occurs across periods in our model, or the governmental review frequency of the possibility to invest in financial institutions in the midst of a crisis. In any case, recalling that the government invests in the equity of banks and other financial institutions, we assume that it will be able to sell the acquired shares to the private sector, after the end of the crisis, for a price that is similar to the purchasing one. In reality, this price is directly linked to the expectation of the future dividends to be paid by the surviving bank. The government could then realise a profit on these investments, after a considerable rise in the aggregate stock market at the end of the crisis, from time $M+1$ onwards (see ref. 45 for a relevant research investigation), or even make a loss. Clearly, any scenario would affect the effective taxpayers loss. In this paper though, we focus solely on the minimisation of taxpayers losses due to bailouts and bank defaults during the crisis episode (time 0 to M), by assuming a neutral realised return on investments in surviving institutions beyond time M.

We define the 4-tuple (S, A_s, P_a, R_a) of the set S of all the states of the dynamic network (namely the state space in which the processes' evolution takes place, leading to all possible configurations of the financial system), set A_s of all actions available to the government from state $s \in S$, transition probabilities $P_a(s,s') = P(s_{t+1}=s' \mid s_t=s, a_t=a)$ between state s at any time t and state s' at time $t+1$ having taken action $a \in A_s$ at time t, and rewards $R_a(s,s')$ (negative losses in our model) received after taking action a at any time t while being at state s and landing in state s' at time $t+1$, where $s,s' \in S$. Furthermore, we consider a constant discount factor γ with $0 \le \gamma < 1$, so that rewards obtained sooner are more relevant. The discounted cumulative reward G from time step t until the end of crisis (recall that a full episode consists of M time steps) is therefore defined by

$$G(t) := \sum_{u=t}^{M-1} \gamma^{u-t} R_{a_u}(s_u, s'_{u+1}). \tag{11}$$

In the remaining of this subsection, we expand on the 4-tuple (S, A_s, P_a, R_a) that defines our MDP and formulate our stochastic control problem.

Firstly, we introduce the MDP states. The states $s_t \in S$, at each time t, in which the financial system may end up, are defined by three main pillars: (a) all the parameters of the network $(W_i(t), E_i(t), PD_i(t), J_i(t), LGD_i, \alpha_i, \mu_i, \sigma_i, w_{ij}, \Sigma_{ij}$, for $i,j \in \{1,\ldots,N\}$, where $w_{ii}=0)$, (b) an indexed set $\mathcal{I}_{def}(t) \subseteq \mathcal{I}$ containing all defaulted nodes prior to time t and (c) the time to maturity $M-t$.

Secondly, we introduce the MDP actions and governmental policies. The MDP actions $a_t \in A_{s_t}$ in our model are the control variables of the government when trying to minimise the losses of the network (i.e. maximise the expected G in (11)). They correspond to injections of capital $a_t \to \boldsymbol{\Delta J}^{\mathbf{a}}(t) := (\Delta J_1^a(t), \Delta J_2^a(t), \ldots, \Delta J_N^a(t))$ increasing the government's investments in the nodes $(1,2,\ldots,N)$, affecting their total wealth and equity according to (4), whose updated (increased) values are denoted by

$$J_i^a(t) := J_i(t) + \Delta J_i^a(t), \quad W_i^a(t) := W_i(t) + \Delta J_i^a(t) \quad \text{and} \quad E_i^a(t) := E_i(t) + \Delta J_i^a(t). \tag{12}$$

These additional resources on one hand, make the nodes more resilient, hence diminishing their updated probability of default PD_i^a via (5), (6), namely

$$PD_i^a(t) := \max\{PDM(W_i^a(t), E_i^a(t), \mu_i, \sigma_i), PDM_i^{floor}\}, \tag{13}$$

leading to (statistically) less defaults due to the updated default mechanism (recall (10)) given by

$$x_i < \Phi_1^{-1}(PD_i^a(t)) \quad \Longleftrightarrow \quad \delta_i^a(t)=1, \tag{14}$$

and consequently to an updated (statistically decreased) total impact

$$I_i^a(t) := \sum_{j\in\mathcal{I}\setminus\{i\}} w_{ij}\delta_j^a(t), \quad \text{for all } i \in \mathcal{I}. \tag{15}$$

On the other hand, these resources will be at risk in case of node i defaulting at time t, since the aforementioned (increased) cumulative investment J_i^a and total wealth W_i^a will both contribute towards an

increased updated taxpayers' overall loss L_i^a given via (7) by

$$L_i^a(t) := \alpha_i W_i^a(t) + LGD_i J_i^a(t). \tag{16}$$

Recalling that each action $a_t \in A_{s_t}$ depends on the current state s_t at any time t, we denote the government policy by a function $\pi(s_t) \to a_t$ that indicates which action to take at each state. A policy that minimises the expected network losses is called optimal policy and is denoted by π_*, while the action a_t^* returned by π_* given a state s_t (i.e. $\pi_*(s_t) \to a_t^*$) is then called the optimal action for that state.

Note that, the model can easily incorporate also the nature of governmental equity injections (asking banks to repay debt, or invest in safer assets to hedge against future losses, or change their strategy in exchange for funding), that would eventually lead to updated $\mu_i^a(t) = \mu_i(\Delta J_i^a(t))$ and $\sigma_i^a(t) = \sigma_i(\Delta J_i^a(t))$ affecting only the resulting updated probability of default $PD_i^a(t)$ in (13), while the rest of our framework and methodology would remain intact.

Thirdly, we introduce the MDP transition probabilities. Within our framework, a node that has defaulted does not contribute to future losses and cannot become active again, i.e. the cardinality of the set of defaulted nodes $|\mathcal{I}_{def}(t)|$ is a non-decreasing function of time t. Hence the transition probability $P_a(s,s')$ from state s to s' will be non-zero only for states s' that: (a) have the same number or more defaulted nodes than state s; (b) are "reachable", in the sense that their characteristics $PD_i(t+1)$, $W_i(t+1)$ and $E_i(t+1)$, for $i \in \mathcal{I} \setminus \mathcal{I}_{def}(t+1)$ (the remaining active nodes in s') take values that are coherent with eqs. (4)–(6) after calculating the impacts $I_i(t)$ from the newly defaulted nodes $i \in \mathcal{I}_{def}(t+1) \setminus \mathcal{I}_{def}(t)$ at time t. For an example on how to identify these so-called reachable states, we refer to the Reachable MDP states example in the Methods section.

Then, for all states s'_{t+1} with a non-zero transition probability $P_{a_t}(s_t, s'_{t+1})$, we can calculate the latter via the Gaussian latent variable model (see also (8), (9)). To be more precise, given the government investments relative to action a_t at state s_t and time t, we use the updated $J_i^a(t), W_i^a(t), E_i^a(t)$ and $PD_i^a(t)$ from (12)–(13) to calculate the transition probability via (see also (8)) the following integral:

$$P_{a_t}(s_t, s'_{t+1}) := \int_D \Phi'_{|\mathcal{I}\setminus\mathcal{I}_{def}(t)|}(\mathbf{u}; \Sigma_{sub}(t))d\mathbf{u}, \tag{17}$$

where Φ' is the density given by (9) with dimension equal to the cardinality of the set of surviving nodes $|\mathcal{I} \setminus \mathcal{I}_{def}(t)| \le N$. Upon recalling the updated version of the default mechanism in (14), the integration domain D in (17) is given by the Cartesian product of the intervals $[-\infty, \Phi_1^{-1}(PD_i^a)]$ for the additional defaulted nodes $i \in \mathcal{I}_{def}(t+1) \setminus \mathcal{I}_{def}(t)$ and the intervals $[\Phi_1^{-1}(PD_i^a), \infty]$ for all the remaining active nodes $i \in \mathcal{I} \setminus \mathcal{I}_{def}(t+1)$ at state s'_{t+1}. The $\Sigma_{sub}(t)$ is the sub-matrix of the original correlation matrix Σ after removing the rows and the columns corresponding to defaulted nodes $i \in \mathcal{I}_{def}(t)$ at state s_t.

Thus, we observe that in our model, the transition probabilities depend exclusively on the government investments a_t, the resulting financial institutions' probability of default PD_i^a and the correlation structure Σ_{ij} with $i,j \in \mathcal{I} \setminus \mathcal{I}_{def}(t)$ which links the financial institutions in the network.

Fourthly, we introduce the MDP rewards. In our model the rewards take non-positive values, since their overall maximisation has to translate for our MDP into the minimisation of the potential overall taxpayers' losses $L_i^a(t)$ in (16) for all nodes $i \in \mathcal{I} \setminus \mathcal{I}_{def}(t)$ after taking action a_t at each time t. Namely, in light of the updated default mechanism (14), we define the reward at time t by

$$R_{a_t}(s_t, s'_{t+1}) := -\sum_{i \in \mathcal{I}\setminus\mathcal{I}_{def}(t)} \left(\alpha_i W_i^a(t) + LGD_i J_i^a(t)\right)\delta_i^a(t), \tag{18}$$

where only the nodes defaulting at time t after taking action a_t, i.e. having $\delta_i^a(t) = 1$, contribute to the sum of losses. This means that the reward at time t can be 0, in case there are no additional defaults occurring at time t.

Finally, we are ready to define the optimal value function and present the stochastic control problem formulation. By doing so, we will formalise the main aim of the government, which is the minimisation of taxpayers' losses during the crisis episode. We therefore need our model to indicate if the government should intervene and if so, which amount it should invest for a given configuration of the financial system to achieve its aforementioned goal. This mathematically translates to the government aiming at finding the optimal actions $a_t^* \in A_{s_t}$, or equivalently the optimal policy π_*, for successive time steps, starting from any time t and any possible state s_t of the dynamic network until the end of the episode at time M, in order to maximise the expected discounted cumulative reward $G(t)$ given by (11).

The optimal value function $V_*(s_t)$ is then defined as the expected discounted cumulative reward $G(t)$ starting from state s_t at time t and following the aforementioned optimal policy π_*, given in light of the definition of rewards (in particular their expression in (18)) by

$$\begin{aligned} V_*(s_t) &:= E_{\pi_*}[G(t)|s_t] = E_{\pi_*}\left[\sum_{u=t}^{M-1} \gamma^{u-t} R_{a_u}(s_u, s'_{u+1}) \middle| s_t\right] \\ &= -E_{\pi_*}\left[\sum_{u=t}^{M-1} \gamma^{u-t} \sum_{i \in \mathcal{I}\setminus\mathcal{I}_{def}(u)} \left(\alpha_i W_i^a(u) + LGD_i J_i^a(u)\right)\delta_i^a(u) \middle| s_t\right] \\ &\forall \quad t \in [0, M-1] \quad \text{and} \quad V_*(s_M) := 0, \end{aligned} \tag{19}$$

where the latter definition follows due to the time step M signifying the end of the crisis episode, when the government can sell all its shares in the banks, thus incurring no additional losses.

Given the definition of π_*, the optimal value function $V_*(s_t)$ represents the maximum expected discounted cumulative reward, which translates into the minimum expected discounted taxpayers' loss, that can be obtained amongst all possible policies π starting from s_t,

$$\begin{aligned} V_*(s_t) &= \max_\pi E_\pi[G(t)|s_t] \\ &= -\min_\pi E_\pi\left[\sum_{u=t}^{M-1} \gamma^{u-t} \sum_{i \in \mathcal{I}\setminus\mathcal{I}_{def}(u)} \left(\alpha_i W_i^a(u) + LGD_i J_i^a(u)\right)\delta_i^a(u) | s_t\right]. \end{aligned} \tag{20}$$

The optimal action value function $Q_*(s_t, a_t)$ is the expected discounted cumulative reward we obtain, if we first take action a_t while being at state s_t and then follow the optimal policy π_* for any of the successive steps from $t+1$ until the end of the episode M. Mathematically, this is defined by

$$\begin{aligned} Q_*(s_t, a_t) &:= E_{\pi_*}[G(t)|s_t, a_t] \\ &= E\left[R_{a_t}(s_t, s'_{t+1})|s_t, a_t\right] + E_{\pi_*}\left[\sum_{u=t+1}^{M-1} \gamma^{u-t} R_{a_u}(s'_u, s'_{u+1})|s_t, a_t\right]. \end{aligned} \tag{21}$$

Similarly to the optimal value function, $Q_*(s_t, a_t)$ represents the maximum expected cumulative reward that can be obtained when starting from s_t and after taking action a_t at time t.

The contribution of the optimal action value function in providing the desired quantitative evaluation required for implementing the model in real-life scenarios is twofold. Firstly, notice that finding Q_* is equivalent to solving the MDP, since the optimal action a_t^* for each state s_t (hence the optimal policy π_*) can be obtained by

$$a_t^* = \underset{a_t}{\text{argmax}}\ Q_*(s_t, a_t). \tag{22}$$

Secondly, we use Q_* in order to quantify the convenience to intervene $\mathrm{Conv}(s_t)$ for the government at each state s_t and any time t, in the forthcoming model implementations. To be more precise, we define by $\mathrm{Conv}(s_t)$ the difference between the optimal action value function corresponding to the best governmental intervention and the optimal action value function associated to a_t^0, which denotes the inaction (no investments) at time t, when being at the state s_t, i.e.

$$\mathrm{Conv}(s_t) := \max_{a_t \in A_{s_t} \setminus \{a_t^0\}} \{Q_*(s_t,a_t)\} - Q_*(s_t,a_t^0). \tag{23}$$

AI technique to solve the MDP

In this subsection, we present our artificial intelligence technique to solve the MDP, driven by our dynamic network of the financial system that can be controlled by a regulator in view of minimising the expected taxpayers' loss.

We firstly recall a standard relationship between optimal value functions and action value functions in MDPs. Observe that the two terms on the right-hand side in the definition (21) of the optimal action value function $Q_*(s_t, a_t)$ are first the immediate expected reward at time t due to taking action a_t and second the optimal expected discounted cumulative reward from time $t+1$ onwards. We can therefore rewrite $Q_*(s_t,a_t)$ from (21) in terms of the transition probabilities (recall (17)) and the future optimal value functions $V_*(s'_{t+1})$ defined in (19), in the form

$$Q_*(s_t,a_t) = \sum_{s'_{t+1}} P_{a_t}(s_t,s'_{t+1})(R_{a_t}(s_t,s'_{t+1}) + \gamma V_*(s'_{t+1})). \tag{24}$$

It is also straightforward to see from the definitions (19) and (21) of the optimal value function $V_*(s_t)$ and action value function $Q_*(s_t,a_t)$, respectively, that

$$V_*(s_t) = \max_{a_t} Q_*(s_t,a_t), \tag{25}$$

i.e. the maximum expected discounted cumulative reward from s_t is the one corresponding to the maximum value of Q_* amongst all available potential actions $a_t \in A_{s_t}$ at time t. Substituting the expression of (24) in (25) thus gives the Bellman optimality equation

$$V_*(s_t) = \max_{a_t \in A_{s_t}} \left\{ \sum_{s'_{t+1}} P_{a_t}(s_t,s'_{t+1}) \left(R_{a_t}(s_t,s'_{t+1}) + \gamma V_*(s'_{t+1}) \right) \right\}. \tag{26}$$

Given that we have a complete description of our MDP (in particular, we have the transition probabilities $P_{a_t}(s_t,s'_{t+1})$ and the rewards $R_{a_t}(s_t,s'_{t+1})$), we could in theory enumerate all possible states, use Dynamic Programming and the Value Iteration algorithm (see ref. 46) to solve our stochastic control problem. This would essentially involve finding V_* using the Bellman optimality equation in (26) and then calculating Q_* via (24), thus solving the MDP. However, applying this standard theory is not a scalable/feasible approach due to (a) the complexity of the MDP states and (b) the enormous number of successor states s' (for all but trivial networks), making standard computations impossible.

We, therefore, propose in this paper an approach to solve the MDP, which involves the use of a variation of the Fitted Value Iteration algorithm (see, e.g. refs. 40, 41) with bespoke characteristics uniquely constructed in our artificial intelligence technique.

Our method consists of the following four steps:

(i) Devise a parametric representation $\overline{V}_*(s,\beta)$ for the optimal value function $V_*(s)$, where β is a placeholder for a set of parameters to fit (see our construction in the Value function approximation subsection, Methods section);

(ii) Use $\overline{V}_*(s,\beta)$ to devise a parametric representation $\overline{Q}_*(s,a,\beta)$ for the optimal action value function $Q_*(s, a)$ in (24) (see the Action value function approximation subsection, Methods section, for its derivation and our technique to calculate it);

(iii) Use $\overline{Q}_*(s,a,\beta)$ for the right-hand side of (25) to obtain an approximate Bellman optimality equation in order to fit β via a learning process (see our technique in the Learning process subsection, Methods section), which will eventually give $V_*(s) \approx \overline{V}_*(s,\beta^{fit})$;

(iv) Finally, use $\overline{V}_*(s,\beta^{fit})$ to calculate $\overline{Q}_*(s,a,\beta^{fit}) \approx Q_*(s,a)$, and hence solve the MDP as in the Optimal solution of the MDP subsection, Methods section.

Each one of the aforementioned steps bares its own difficulties and technical obstacles, which we overcome in the analysis presented in the aforementioned subsections of the Methods section.

In the following subsections, we use our artificial intelligence technique to solve the MDP in two implementation case studies. We show how our model works and obtain qualitative results on the optimal bailout decision problem faced by governments. A professional calibration of our model would require the effort and firepower of a central bank or a government office, and access to sensitive data. Nonetheless, by exploring these two case studies, we provide useful insights for whether and when taxpayers should fund bank bailouts.

Setup of implementation case studies

To illustrate how our method works and its potential, we apply our model on both a synthetic homogeneous network (Krackhardt kite graph) and a real network of the European global systemically important institutions. Before we present our case studies, we assign values to a set of parameters, that are common to both case studies (unless otherwise specified). We consider a crisis episode that will last for $M = 7$ time steps, a discount factor $\gamma = 0.98$ and an initial government investment $J_i(0) = 0$ for each node (bank or other financial institution) $i \in \mathcal{I}$. Moreover, we assume the percentages α_i of wealth loss upon default of node i to be all the same, i.e. $\alpha_i = \alpha$, and we conservatively assume that the expected value of the total wealth's return is $\mu_i = 0$, for all $i \in \mathcal{I}$. Recall that, we are considering equity investments by the government that can be recovered, in case of default, only after all the depositors and bond holders are satisfied. Hence, we assume that the government loses all its investments in case of default, i.e. $LGD_i = 1$, for all $i \in \mathcal{I}$. If the government is allowed to use other means besides equity investments, e.g. bond investments, then $LGD_i \in (0, 1)$ (see ref. 47 for a study on senior and subordinated recovery rates). However, this would imply a softer effect on the solvency issues and require a modification of the probability of default formula in (13), since the investment would not be directly affecting the equity (12) anymore. In order to take into account the average correlation between financial institutions, we use a homogeneous correlation matrix for our nodes, which is set to $\Sigma_{ij} = 0.5$ for $i \neq j \in \mathcal{I} \setminus \mathcal{I}_{def}$ following ref. 48. The volatility σ_i of the total wealth's return for each $i \in \mathcal{I}$, is calculated at time 0 by inverting (5) using the initial (known) values of $PD_i(0)$, $E_i(0)$ and $W_i(0)$. The values of σ_i, $i \in \mathcal{I}$, are then assumed to remain constant at successive time steps of the simulation, from time 1 to M. Finally, we set the floor of the probability of default for each node i as $PDM_i^{floor} = 0.00021$, which is the upper end of the AAA default probability bracket within the internal credit rating methodology used by Credit Suisse[49]. In the sequel, we denote the available governmental investment actions by

<node> @ <capital investment as a tenth of a percent of the total asset W>,

with the convention that an action that considers all nodes is indicated with <node> = 0. For example, 8@05 means an investment of 50 *bp*

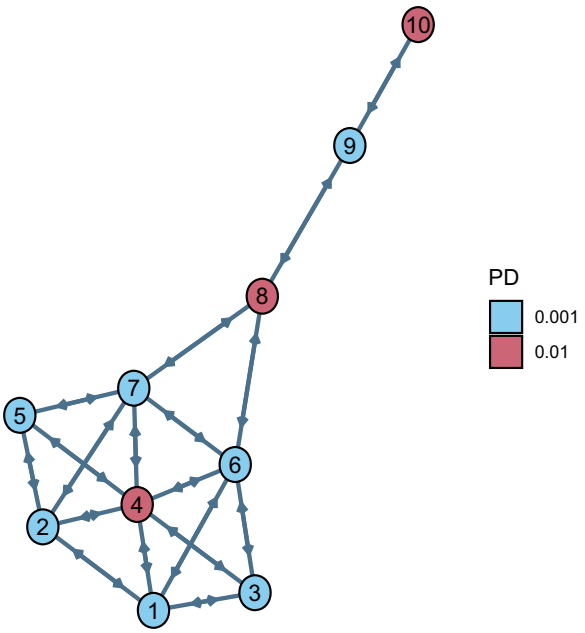

Fig. 1 | Krackhardt Kite (KK) network. The KK network is used to assess how bailout decisions are influenced by node centrality. The ten nodes of the graph in $\mathcal{I} = \{1, \ldots, 10\}$ represent financial institutions, which are identical apart from their probabilities of default (PD) at time 0, which are $PD_i(0) = 0.01$, for $i \in \{4, 8, 10\}$, and $PD_i(0) = 0.001$, for $i \in \mathcal{I} \setminus \{4,8,10\}$. They all have a normalised total asset $W_i(0) = 100$ and capital $E_i(0) = 3$. The edges between nodes, representing claims between financial institutions, are oriented and homogeneous, assuming the value $w_{ij} = 1$, for all $i \neq j \in \mathcal{I}$.

W_8 or 0.5% W_8 in node 8, while 0@15 stands for an investment of 1.5% W_i in each node $i \in \mathcal{I} \setminus \mathcal{I}_{def}$.

The common theme is that adding external resources makes the network more resilient, but such resources can be lost in a subsequent default, which creates a trade-off for the decision-maker. The optimal policy that balances this trade-off and minimises the overall expected taxpayers' loss is an optimal solution to our MDP, and is analysed in the forthcoming two studies.

Case study 1: KK network

This study concerns a network with homogeneous nodes organised as the Krackhardt kite (KK) graph (Fig. 1, see also ref. 50), which is referred to as the KK network. The main reason for choosing the KK graph as an underlying network is to primarily assess whether our algorithm can distinguish between central nodes and peripheral ones. In particular, we use the network characteristics in terms of the centrality of nodes 4, 8 and 10 (see Fig. 1), to investigate how bailout decisions depend on the nodes' position in the network. However, we will also investigate additional hypotheses and reach important financial conclusions on bailout decision making.

In this case study, all the nodes (banks or other financial institutions) have total asset $W_i(0) = 100$ and capital $E_i(0) = 3$. As shown in Fig. 1, the nodes in red colour have probability of default $PD_i(0) = 0.01$, for $i \in \{4, 8, 10\}$, while the others have $PD_i(0) = 0.001$, for $i \in \mathcal{I} \setminus \{4,8,10\}$. The edges between nodes are oriented and homogeneous, assuming the value $w_{ij} = 1$, for all $i \neq j \in \mathcal{I}$. For the sake of this case study, we restrict the potential investment amounts in each node i to be: 0, 0.5% W_i, 1% W_i, 1.5% W_i or 2% W_i. Furthermore, the government can choose at each time step to invest in the single nodes 4, 8, 10 or in all nodes.

The optimal action value function $Q^*(s_0, a_0)$ at time 0 is illustrated in Fig. 2 for three scenarios of percentages of wealth loss upon default. In particular, for a "relatively low" $\alpha = 0.0001$, the best action (minimising losses) is not to invest in any bank (0@0). Moving from the top to the bottom panel (as α increases) the option not to invest becomes more and more costly to the system. For a "relatively intermediate" $\alpha = 0.001$, not investing is roughly equally favourable to investments in single financial institutions, while for a "relatively high" $\alpha = 0.01$, the best action becomes to invest 1.5% W_i in all financial institutions (0@15). It is also interesting to note that (see Fig. 2) irrespective of the α-value: (i) investing the maximum amount of 2% W_i in all banks (0@20) is never the best choice; (ii) providing the minimum capital (0@05) is always the worst choice, as the additional investment is too small to make them resilient, but still increases the funds at risk in case of default. The sensitivity of the optimal policy with respect to α will be further examined in more detail also in our next (more realistic) case study.

Fixing the percentage of wealth loss upon default at $\alpha = 0.0001$, we now focus our analysis on the central node 4, representing a financial institution with multiple links and interconnections with its peers, versus the peripheral node 10, representing a relatively isolated financial institution linked only with one other (see Fig. 1). The results in Fig. 3a conclude that investing in the central node 4 is always better than in the peripheral node 10 for the same amount of capital and all such choices. Thus, our algorithm indeed shows a clear preference in central rather than peripheral node investments.

However, the results change when the government had already invested even the minimum possible amount of 0.5% W_{10} in the peripheral node 10. In this case, Fig. 3b with $J_{10}(0) = 0.5$ concludes that a substantial additional investment in bank 10 (namely, 10@15 or 10@20) largely outperforms any other strategy—including not investing at all, and all types of investments in the central node 4. Such a result indicates that the optimal strategy for the government is therefore to keep investing (sufficiently high amounts of capital) in node 10, aiming at saving this already invested capital $J_{10}(0) = 0.5$. This governmental tendency to provide capital to distressed banks if they had already invested in them creates moral hazard, as the bank could act haphazardly relying on the implicit government guarantee. The fact that bailouts create moral hazard has been emphasised extensively in the financial and economic literature, both theoretically and empirically (see e.g., refs. 51–54). Moreover, given that the assumed $J_{10}(0) = 0.5$ is the worst amongst all possible investments in a single node at time 0 (see both Fig. 2 and Fig. 3a), the suggested optimal, additional, significantly large investment in the peripheral node 10, could be also viewed as an eventual strengthening of the originally weak investment of 0.5% W_{10}.

Lastly, we perform a sensitivity analysis of the optimal action versus the duration of crisis. Irrespective of the type of investment, the results in Fig. 3a show that the optimal action value function Q^* decreases in absolute value (i.e. size of losses decreases) as the time to the end of the episode $M-t$ decreases. The main reason is that each node of the network is unstable with an associated probability of default per unit time, hence the shorter the time horizon the lower the expected losses. Furthermore, the contagion has less time to propagate, which explains also why the node's position in the network becomes less and less relevant.

Case study 2: EBA network

After having considered a small synthetic graph in the first case study, we now study a network of the European global systemically important institutions (GSII, see Table 1) obtained from the data provided by the European Banking Authority[55], which is referred to as the EBA network. Note that, the original data do not contain the complete bilateral network of exposures, as this is considered business-sensitive information. While the specific exposure between two banks is unknown, the aggregated credit exposure of a bank versus other financial institutions is provided. For each bank i in the set $\mathcal{I}$ of the European GSII, we have its total inter-bank asset $\sum_{j \in \mathcal{I}} w_{ij}$ and liability $\sum_{j \in \mathcal{I}} w_{ji}$, which

https://doi.org/10.1038/s41467-022-34102-1

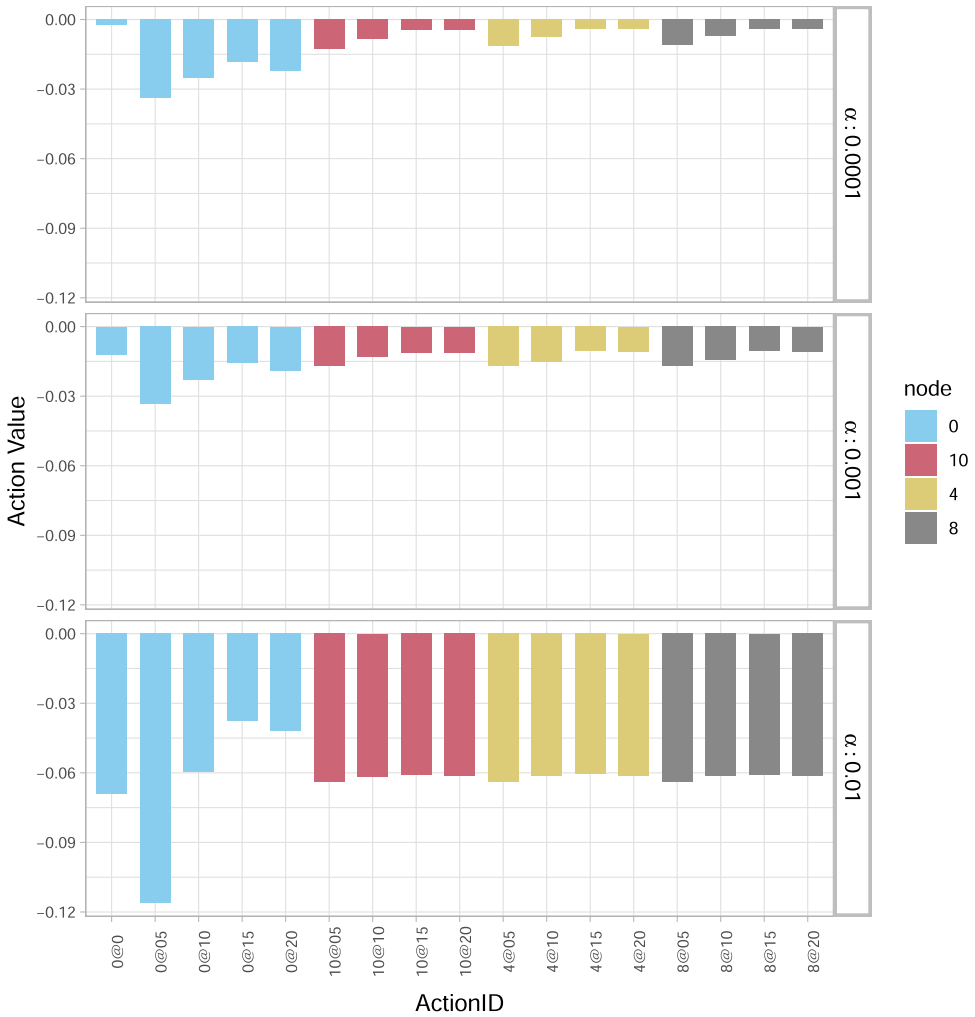

Fig. 2 | Optimal action value function Q_* for the Krackhardt Kite (KK) network. The optimal action value function $Q_*(s_0, a_0)$ at time 0 for different actions a_0 (a government investment of 0, 0.5, 1, 1.5 or 2 in the equity of the nodes) and values of percentage wealth loss upon default α (0.0001, 0.001 and 0.01). In the legend, the colours identify the nodes {0, 10, 4, 8} in the figure (0 represents all nodes). On the x axis, the ActionID 0@0 means no investment, 0@05 means investing 0.5 in all the nodes, 10@05 means investing 0.5 in node 10, etc. For a small value of $\alpha = 0.0001$, the best action is not to invest (0@0). As α increases, so does the convenience of investing more capital. For $\alpha = 0.01$ the best action (corresponding to smallest loss) is to invest 1.5 in all the nodes (0@15). It is never convenient to invest the maximum amount of capital (0@20), while investing 0.5 in all the nodes (0@05) is the worst action for all values of α, as the additional capital is not enough to strengthen the network and it is at risk following potential defaults.

can be used to reconstruct a network that satisfies the constrains (see, e.g. algorithms described in refs. 37, 42). The reconstructed network (see Fig. 4) can be different but has similar characteristics to the actual network of bilateral exposures. The values of total asset $W_i(0)$ and capital $E_i(0)$ at time 0 used for each financial institution i are reported in Table 1. The probabilities of default are derived using data from the credit rating agency Fitch[56] and show that the nodes with the higher probability of default are Monte dei Paschi di Siena (MPS) and BFA (see both Table 1 and Fig. 4).

To facilitate our analysis, we firstly pretend that the European Union (including the UK) is a fiscal union with a single regulator ("government") that is accountable to all European taxpayers. Then, we consider any individual states' investments in banks prior to 2014 as "private" investments, hence we set the initial regulator investments to be $J_i(0) = 0$ for all $i \in \mathcal{I}$. For the sake of this case study, we restrict the potential investment amounts to inject in each financial institution i to be: 0, 0.5% W_i, 1% W_i, 1.5% W_i, 2% W_i, 2.5% W_i or 3% W_i. Furthermore, we assume that the government can choose at each time step t to invest in all the nodes that are considered "risky" at that time, defined as each financial institution $i \in \mathcal{I} \setminus \mathcal{I}_{def}$ with $PD_i > 0.009$, according to our (arbitrarily) chosen threshold. In this case study, the notation 0@05 thus indicates an investment of 0.5% W_i in each risky node $i \in \mathcal{I} \setminus \mathcal{I}_{def}$.

For a detailed quantitative and qualitative analysis, we rely on the convenience measure Conv for the government, defined in (23), to intervene with equity investments and we analyse the system for four different percentages $\alpha \in \{0.0001, 0.001, 0.005, 0.01\}$ of wealth loss upon default. We observe from Fig. 5a that, we have a convenience Conv > 0 for higher percentages of wealth loss upon default ($\alpha = 0.01$, 0.005 and 0.001), thus investing is a favourable action, while Conv < 0 for smaller α ($\alpha = 0.0001$), implying that it is not convenient for the government to invest. This is consistent with our first case study using the KK network (previous subsection), as investing an amount of capital is convenient only for relatively high values of α, in order to make the network sufficiently resilient.

A sensitivity analysis of the convenience to intervene for the government is further examined versus the duration of crisis, the initial capital and the credit exposures of financial institutions:

(a) The results in Fig. 5a further conclude that the convenience to intervene tends to be an increasing function of the time to the end

https://doi.org/10.1038/s41467-022-34102-1

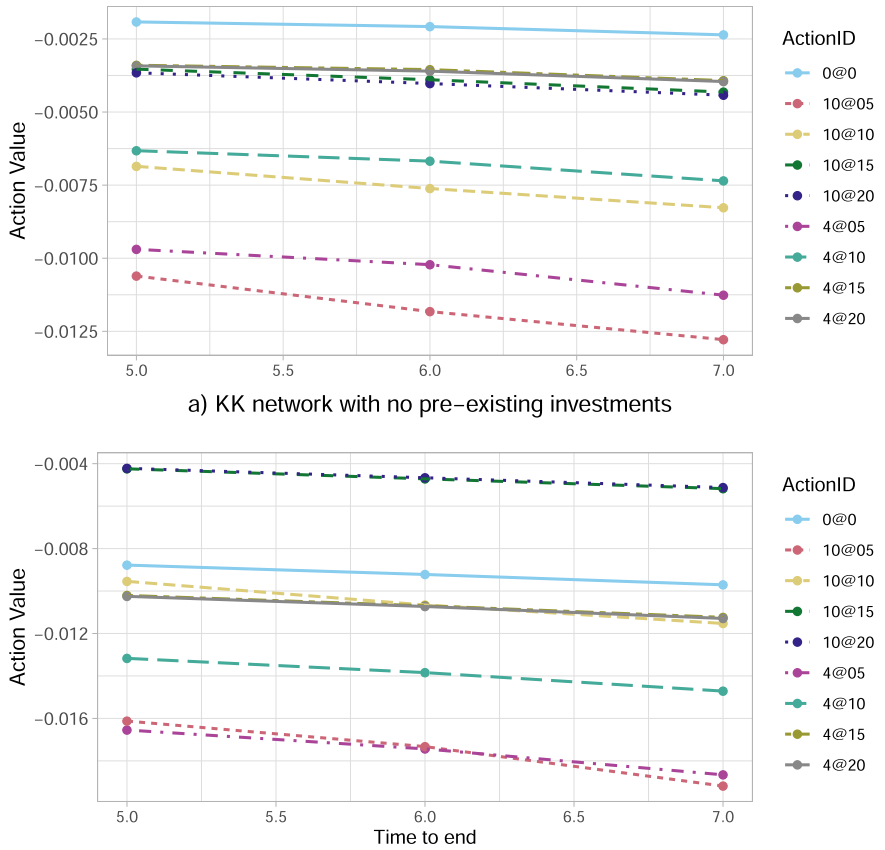

Fig. 3 | Optimal action value function Q_* for the Krackhardt Kite (KK) network as a function of time to the end of crisis. The results are obtained for the percentage $\alpha = 0.0001$ of wealth loss upon default, and focus on nodes 4 (central node) and 10 (peripheral node). In the legend, the ActionID 0@0 means no investment, 0@05 means investing 0.5 in all the nodes, 10@05 means investing 0.5 in node 10, etc. **a** The algorithm feels the network structure and suggests to invest in node 4 (leading to smaller loss) rather than node 10. **b** In case the government had previously invested in node 10, the government needs to protect its investment by optimally risking an additional investment in node 10.

of the episode $M-t$ when Conv > 0 and a decreasing function when Conv < 0. This implies that the convenience to intervene or not, weakens (decreases in absolute value) as we approach the end of the crisis. Interestingly though, it appears that the nature of the action does not change with time, since the function Conv does not change sign.

(b) Our results in Fig. 5b further show that the convenience to intervene Conv is dependent on the banks' resilience, expressed via the initial capital $E_i(0)$ of each bank i. In particular, this severely distressed version of the network, where the value of $E_i(0)$ has been artificially halved compared with the original case study presented in Fig. 5a, has the effect of increasing the convenience Conv for the government to intervene for each value of α. A more thorough analysis in Fig. 6a further reveals that the convenience to intervene increases on average as the financial institutions' initial capital $E_i(0)$ decreases, for all duration lengths of the crisis. It is interesting to also note that, this convenience intensifies significantly for larger lengths of time until the end of the crisis.

(c) It is also clear from the results in Fig. 6b that the convenience to intervene increases as the bilateral credit exposures w_{ij} between financial institutions across the whole network increase. It is interesting to further observe that the impact of longer crisis duration on the convenience to intervene is massive.

A sensitivity analysis of the governmental optimal action is also examined versus the discount factor, the financial institutions' probabilities of default and credit exposures, the percentage of wealth loss upon default and the initial capital:

(a) Our analysis in Fig. 7a shows clearly that the optimal action value function $Q_*(s_0, a_0)$ at time 0 decreases for all potential actions as the discount factor γ increases. That is, as the future losses become more relevant (from the governmental point of view), the expected systemic losses increase in absolute value, while the optimal action does not change qualitatively.

(b) Our results in Fig. 7b then show that the optimal action value function $Q_*(s_0,a_0)$ at time 0 decreases (taxpayers' losses increase in absolute value) for all potential actions, with increasing probabilities of default. Two interesting features appearing are: (i) the initially narrowly optimal action (0@30) becomes clearly optimal as the probabilities of default increase; (ii) the worst possible action, namely the one to avoid, changes from the smallest possible investment of $0.5\%W_i$ to larger investments of $1.0\%W_i$, $1.5\%W_i$ in all risky financial institutions i. That is, even these medium size additional investments are not enough to make them resilient, but still significantly increase the funds at risk in case of default.

(c) Our results in Fig. 7c also show that the optimal action value function $Q_*(s_0, a_0)$ at time 0 decreases for all potential actions, with increasing bilateral credit exposures w_{ij} between financial institutions across the whole network. We also note that the difference in the performance of the optimal investment of a large amount (0@30) and non investing at all, increases with greater credit exposures amongst institutions.

Table 1 | European Union's Global Systemically Important Institutions (GSII)

SYMBOL	W	E	PD	BANK
BFA	235	12	0.0116	BFA
MPS	201	7	0.0093	Monte dei Paschi di Siena
UNI	1034	45	0.0017	Unicredit
INT	696	38	0.0017	Intesa Sanpaolo
CAI	377	19	0.0017	La Caixa
BNP	2253	70	0.001	BNP Paribas
BAR	1940	59	0.001	Barclays
CAG	1723	71	0.001	Credit Agricole
DEB	1659	63	0.001	Deutsche Bank
SAN	1456	64	0.001	Santander
RBS	1411	51	0.001	RBS
SOC	1409	47	0.001	Societe Generale
BPC	1337	50	0.001	BPCE
ING	1164	41	0.001	ING
LOY	1107	46	0.001	Lloyds
BBV	723	42	0.001	BBVA
CMU	695	37	0.001	Credit Mutuel
COM	656	25	0.001	Commerzbank
DAN	494	19	0.001	Danske Bank
ABN	421	16	0.001	ABN Amro
DZB	356	13	0.001	DZ Bank
DNB	332	15	0.001	DNB
SEB	310	13	0.001	SEB
LBW	290	13	0.001	LBBW
BLB	275	10	0.001	Bayern LB
SWE	249	10	0.001	Swedbank
KBC	232	14	0.001	KBC
POS	223	7	0.001	Banque Postale
ERS	219	11	0.001	Erste Group
NLB	216	7	0.001	NordLB
HLB	199	8	0.001	Helaba
HSB	2680	117	0.0004	HSBC
RAB	728	34	0.0004	Rabobank
NOR	655	25	0.0004	Nordea
HAN	334	11	0.0004	Handelsbanken

The total asset (W) and Tier 1 capital (E) are expressed in billions of EUR. The data are from the European Banking Authority (EBA)[55] and are relative to the end of 2014. The probabilities of default have been derived using data from the Fitch credit rating agency[56].

(d) It has already been confirmed by both case studies under consideration (KK and EBA network subsections), that as the percentage α of potential wealth loss upon default increases, the inaction (no investments) becomes less convenient for the government. We now aim to explore further the transition between the scenarios when it is convenient and not convenient for a regulator to intervene, by studying the optimal action values $Q_*(s_0, a_0)$ at time 0 with respect to changes in α. Our results in Fig. 8a conclude that: (i) there exists a critical $\alpha_c \approx 0.00079$ that splits the parameter space of α-values into two "wealth loss regimes" of high/low values, reflecting governmental action/inaction, respectively; (ii) the optimal action at time 0 changes drastically (non-smoothly) from a do not invest anything policy for α just below α_c to an invest the maximum amount of $3.0\%W_i$ in all risky institutions i (0@30) policy for α just above α_c. Notice that, these actions are in fact the two extremes. This is an interesting result as one might have expected a smoother transition between optimal actions as α increases.

(e) We then conclude from our results in Fig. 8b, which is a severely distressed version of the original network (presented in Fig. 8a), where the financial institution i's capital $E_i(0)$ has been artificially halved, that the optimal action value function $Q_*(s_0, a_0)$ at time 0 changes significantly both quantitatively and qualitatively. Interestingly, the optimal action for $\alpha > \alpha_c$ becomes the considerably decreased investment of $1.5\%\ W_i$ in all risky financial institutions i (0@15), compared to the original network (see Fig. 8a). A more thorough analysis in Fig. 9 reveals that, as the initial capital $E_i(0)$ decreases, the optimal investment amount indeed decreases as well. In particular, we observe that the investment of $3.0\%W_i$ in all risky financial institutions i, becomes $2.5\%W_i$ when the capital decreases by 25% and $1.5\%W_i$ when the capital decreases by 50%. Furthermore, our results in Fig. 9 reveal that the universally (for all $E_i(0)$) worst action is to invest the lower amount of $0.5\%W_i$ in all risky financial institutions i, as such an additional investment is too small to make them resilient, but still increases the funds at risk in case of default.

Lastly, we perform a sensitivity analysis of the aforementioned wealth loss regimes versus the financial institutions' initial capital. We can conclude from our results in Fig. 8b that, as the financial institution i's capital $E_i(0)$ decreases, the value α_c separating the two wealth loss regimes of unfavourable and favourable regulatory interventions, is significantly lower $\alpha_c \approx 0.0005$. Namely, the government is willing to intervene for much lower percentages of wealth loss.

Discussion

The main theoretical contribution of this paper is twofold. On the modelling side, we propose a framework with a dynamic network of financial institutions, which allows governments or regulatory bodies to assess and quantify bank bailout decisions; we further show how these decisions can be cast into actions in a Markov Decision Process (MDP), where the states of the MDP are defined in terms of the underlying network of financial exposures and the MDP dynamics is derived from the network dynamics. On the optimisation side, in order to identify the optimal governmental investment policy from the taxpayers' standpoint, for each state of the financial institutions network at each time t, we develop a methodology involving artificial intelligence techniques that learn the optimal bailout actions to minimise the taxpayers' losses (for details refer to the Methods section). This methodology goes beyond standard theory and overcomes several technical hurdles, which could also have a significant impact on tackling stochastic control problems in dynamic networks with few reward signals.

In the implementation of our framework and methodology in two case studies, it is evident from our analysis that the loss for the taxpayers, as a fraction α of the financial institution's total assets upon default, plays a central role in systemic risk modelling. In our main case study, which uses the data relative to the European global systemically important institutions, we find that governmental interventions do not improve the expected loss of the financial network if $\alpha < \alpha_c$, for some critical threshold value α_c, which is determined by the network's characteristics and is decreasing as the distress of the network increases. We also find that the convenience to intervene increases for longer crisis time horizons, smaller banks' resilience (equity) and higher bank bilateral credit exposures. Moreover, using a well-known small graph (the Krackhardt kite network) as a simplified case study, we find that even though investing in central nodes is a priori more favourable, the government should optimally keep investing in a node if it had already invested in it in the past, even if that node is not a central one in the network. The government needs to evaluate carefully a potential investment, since the rescued financial institution could increase its risky investments knowing that it

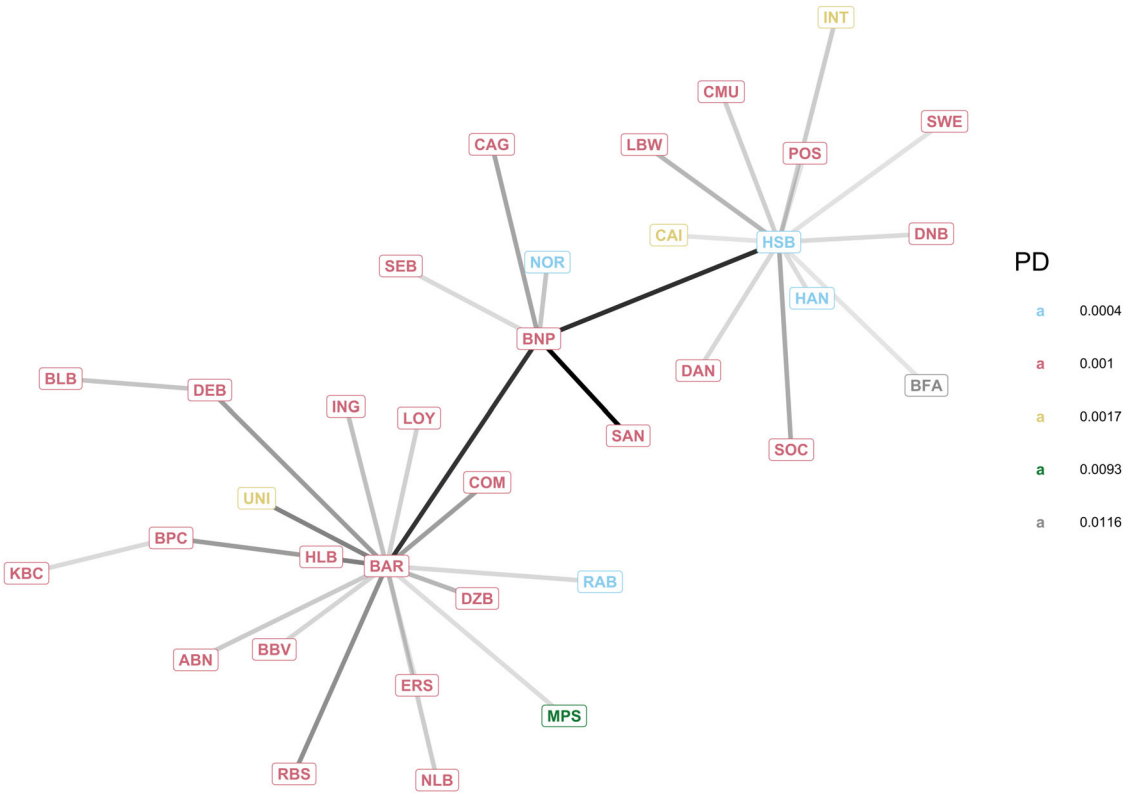

Fig. 4 | Maximum spanning tree of the European Banking Authority (EBA) network. The network of the European Union's Global Systemically Important Institutions (GSII) has been reconstructed from aggregated data available at the EBA website[55]. Each node represents a financial institution (see Table 1), its colour represents its probability of default (PD), and darker edges identify stronger exposures.

would be bailed-out in case it became distressed again, thus leading to moral hazard.

The modelling and optimisation methods in this paper can be extended in several directions. The decision maker could introduce a second layer of stochasticity to the model, driving the market conditions after the end of the crisis. This would allow the additional optimisation of the sale timing and price of the acquired shares of rescued (surviving) financial institutions, in view of maximising the taxpayers' profits or minimising their losses from these investments. Another possible extension could be the consideration of an action-dependent end of the crisis time $M_t := M(J_1(t),\ldots,J_N(t))$, whereby large bailout decisions could shorten the duration of the crisis. Such extensions would require the use of other methodologies, which could lead to interesting future research projects.

Methods

Value function approximation

In order to solve our MDP using our variation of a Fitted Value Iteration algorithm, we need a parametric representation of the optimal value function $V_*(s_t)$. In our case, we see from (20) that $V_*(s_t)$ is minus the minimum expected cumulative loss when starting from state s_t (i.e. the maximum expected discounted cumulative reward from state s_t), incurred between time t and the end of the episode at time M.

The potential for additional losses in the financial network increases with the number of (surviving) nodes and number of residual steps $m := M-t$ until the end of the crisis. It is thus natural to try to approximate $V_*(s_t)$ by a (weighted) sum of the expected direct loss contributions due to each individual node at each of the remaining m time steps (cf. optimisation criterion in (20)).

We estimate these expectations using an approximation of each node's probability of default and overall wealth at all remaining time steps. To do so, our methodology prescribes to firstly use the current probabilities of default for each step to approximate the expected future impact of defaults on each node i at the next time step. These expected impacts are then used as input in order to estimate the (expected) future wealth, equity and obtain an updated approximate probability of default for each node i at that next time step, which closes the loop. By following this procedure, we are then able to obtain an approximation $Z_{ik}(s_t)$ of the expected direct loss contribution of each individual node $i \in \mathcal{I} \setminus \mathcal{I}_{def}(t)$ at each remaining time step $k \in \{1,\ldots,m\}$, for any arbitrary policy π.

In order to obtain the matrix $\overline{\mathbf{Z}} := (\overline{Z}_{ik}(s_t))$ that will be used to define our value function approximation, we take into account potential government interventions to minimise the taxpayers' losses. This involves a sequential optimisation whereby, at each remaining step k, the government chooses the action $\overline{a}_{t+k-1}$ that minimises the sum of the contributions Z_{ik}, assuming no intervention at future steps and using the already chosen actions for previous steps. Further mathematical details of our choice of $\overline{\mathbf{Z}}$ in terms of the characteristics of the network are given in the following Value function parametrisation subsection.

Our ansatz for the parametric representation $\overline{V}_*(s_t,\beta)$ of $V_*(s_t)$ is that it is given by a linear combination of the elements $\overline{Z}_{ik}$, in which the coefficients (weights) β are arranged in a matrix that can change with time, i.e. $\beta \equiv \beta_t := (\beta_{ik}(t))$. Namely,

$$\overline{V}_*(s_t,\beta_t) := -\sum_{i,k} \beta_{ik}(t)\overline{Z}_{ik}(s_t), \qquad \text{for } i \in \mathcal{I} \setminus \mathcal{I}_{def}(t), k \in \{1,\ldots,m\}. \tag{27}$$

https://doi.org/10.1038/s41467-022-34102-1

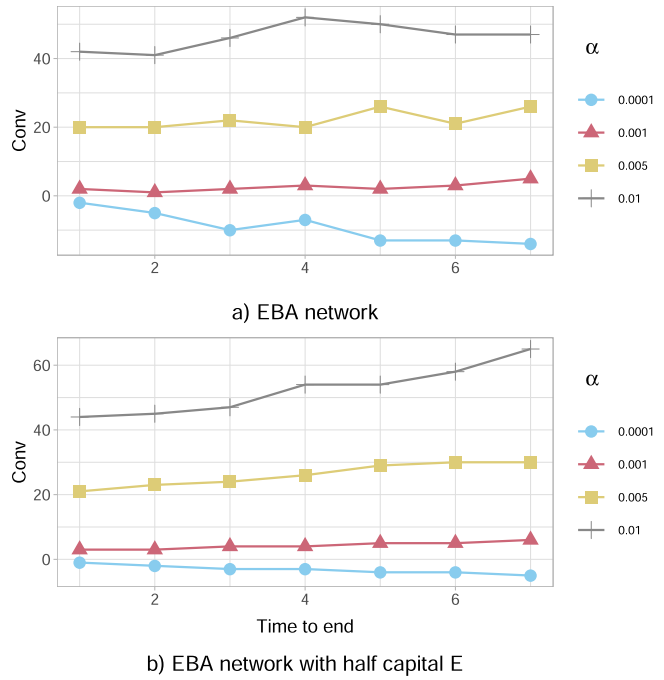

Fig. 5 | The convenience to intervene Conv for the European Banking Authority (EBA) network as a function of time to the end of crisis. a The Conv (in millions of EUR) defined in (23) is positive for higher percentages of wealth loss upon default ($\alpha = 0.01, 0.005, 0.001$), thus investing is a favourable action. Conversely, Conv < 0 for smaller α ($\alpha = 0.0001$), implying that it is not convenient for the government to invest. Conv tends to be an increasing function of the time to the end of the crisis when positive, and a decreasing function when negative. **b** A severely distressed version of the network, where the banks' capital $E_i(0)$ has been artificially halved (all other characteristics are the same). We observe that such a distress has the effect of increasing Conv for each value of α.

Value function parametrisation

In this subsection, we detail our choice for $(\overline{Z}_{ik}(s_t))$ used in our ansatz for the value function approximation $\overline{V}_*(s_t,\beta)$ in (27), with node $i \in \mathcal{I} \setminus \mathcal{I}_{def}(t)$ and artificial step $k \in \{1,\ldots,m = M-t\}$, where each k corresponds to the time step $t+k-1$. The last step $k=m$, therefore, refers to time $M-1$, namely the last step away from the end of the episode. In the following, we denote by W_i, E_i and J_i the current levels of wealth, equity and cumulative investment at time t.

We introduce the auxiliary matrix $\mathbf{Z}$, with elements $Z_{ik}(s_t;a_t,\ldots,a_{t+k-1})$ representing our approximation of the expected direct loss contribution due to the default of node i at time $t+k-1$, taking into account the government actions a_{t+j-1} at times $t+j-1$, for all $j=1,\ldots,k$. That is, we approximate the expected values of the summands involved in the optimisation criterion (20) via the terms

$$Z_{ik} := \begin{cases} PD_{ik} L_{ik}, & \text{if } k=1; \\ PD_{ik} \gamma^{k-1} L_{ik} \prod_{r=1}^{k-1}(1-PD_{ir}), & \text{if } k>1, \end{cases}$$

where, the approximate probability of default of each node i at time $t+k-1$ is given by a value PD_{ik} and the approximate taxpayers' loss by a value L_{ik}. Note that in this approximation, for a step $k>1$, a node i can contribute to the expected loss, only if it has not defaulted in the previous steps; hence, the presence of the survival probabilities $1-PD_{ir}$, for all $r \in \{1,\ldots,k-1\}$.

To be more precise with the above approximations, we begin by approximating each node's probability of default (cf. its original definition in (13)) via

$$PD_{ik} := \max\left\{PDM(W_i + J_{ik} - I_{ik}, E_i + J_{ik} - I_{ik}, \mu_i, \sigma_i), PDM_i^{floor}\right\},$$

which takes into account the potential cumulative investment J_{ik} from the government and our approximation of the expected cumulative impact I_{ik} on node i up to time $t+k-1$. On one hand, the cumulative investment J_{ik} in node i is a function of the actions $(a_t,\ldots,a_{t+k-1})$ that the government can take between t and $t+k-1$ and is independently defined (via its standard definition) by

$$J_{ik}(a_t,\ldots,a_{t+k-1}) := \sum_{r=1}^{k} \Delta J_i^a(t+r-1).$$

On the other hand, we approximate the expected cumulative impact *of* I_{ik} via the approximated probability of defaults of all nodes $j \in H := \mathcal{I} \setminus (\mathcal{I}_{def}(t) \cup \{i\})$ at the previous time steps, which creates the desired approximation loop. Namely, (cf. its original definition in (15)) we obtain the approximation

$$I_{ik} := \begin{cases} 0 & \text{if } k=1; \\ \sum_{j\in H} PD_{j1} w_{ij} & \text{if } k=2; \\ I_{ik-1} + \sum_{j\in H} PD_{jk-1} w_{ij} \prod_{r=1}^{k-2}(1-PD_{jr}) & \text{if } k>2. \end{cases}$$

Consequently, I_{ik} is used as an input to approximate the taxpayers' loss L_{ik} due to the default of node i at time $t+k-1$ (cf. its original definition in (16)) via

$$L_{ik} := \alpha_i(W_i + J_{ik} - I_{ik}) + LGD_i(J_i + J_{ik}).$$

At this point, we note that each $Z_{ik} = Z_{ik}(s_t;a_t,\ldots,a_{t+k-1})$ depends on the actions $(a_t,\ldots,a_{t+k-1})$ via the terms $J_{ik}(a_t,\ldots,a_{t+k-1})$ involved in both PD_{ik} and L_{ik}. The only remaining task is therefore to provide an approximation for the actions $(a_t,\ldots,a_{t+k-1})$ that optimise the aforementioned quantities Z_{ik} according to the objective in our stochastic control problem defined in (20). To that end, we define the total expected direct loss contribution $TL(s_t;a_t,..,a_{M-1})$ as an approximation of the optimisation criterion in (20) for any arbitrary policy π, or equivalently any actions $a_t,..,a_{M-1}$. Namely, we aim at finding an approximation $(\overline{a}_t,\ldots,\overline{a}_{M-1})$ for the actions that minimise (over all nodes $i \in \mathcal{I} \setminus \mathcal{I}_{def}(t)$ in any of the remaining time steps $k \in \{1,\ldots,m\}$)

$$TL(s_t; a_t, a_{t+1},..,a_{M-1}) := \sum_{i,k} Z_{ik}(s_t; a_t, a_{t+1},..,a_{t+k-1}).$$

To do so, we calculate each $\overline{a}_{t+j-1}$ sequentially for each step $j \in \{1,\ldots,m\}$ as follows:

$$\begin{aligned} \overline{a}_t &:= \operatorname*{argmin}_{a_t} TL(s_t; a_t, a_{t+1}^0, a_{t+2}^0,..,a_{M-1}^0) \\ \overline{a}_{t+1} &:= \operatorname*{argmin}_{a_{t+1}} TL(s_t; \overline{a}_t, a_{t+1}, a_{t+2}^0,..,a_{M-1}^0) \\ &\vdots \\ \overline{a}_{M-1} &:= \operatorname*{argmin}_{a_{M-1}} TL(s_t; \overline{a}_t, \overline{a}_{t+1}, \ldots, \overline{a}_{M-2}, a_{M-1}), \end{aligned}$$

with a^0 denoting the action corresponding to no additional government investment. Then, the specific matrix $\overline{Z} = (\overline{Z}_{ik}(s_t))$ involved in our value function approximation $\overline{V}_*(s_t,\beta)$ in (27) is defined by

$$\overline{Z}_{ik}(s_t) := Z_{ik}(s_t; \overline{a}_t, \ldots, \overline{a}_{t+k-1}).$$

https://doi.org/10.1038/s41467-022-34102-1

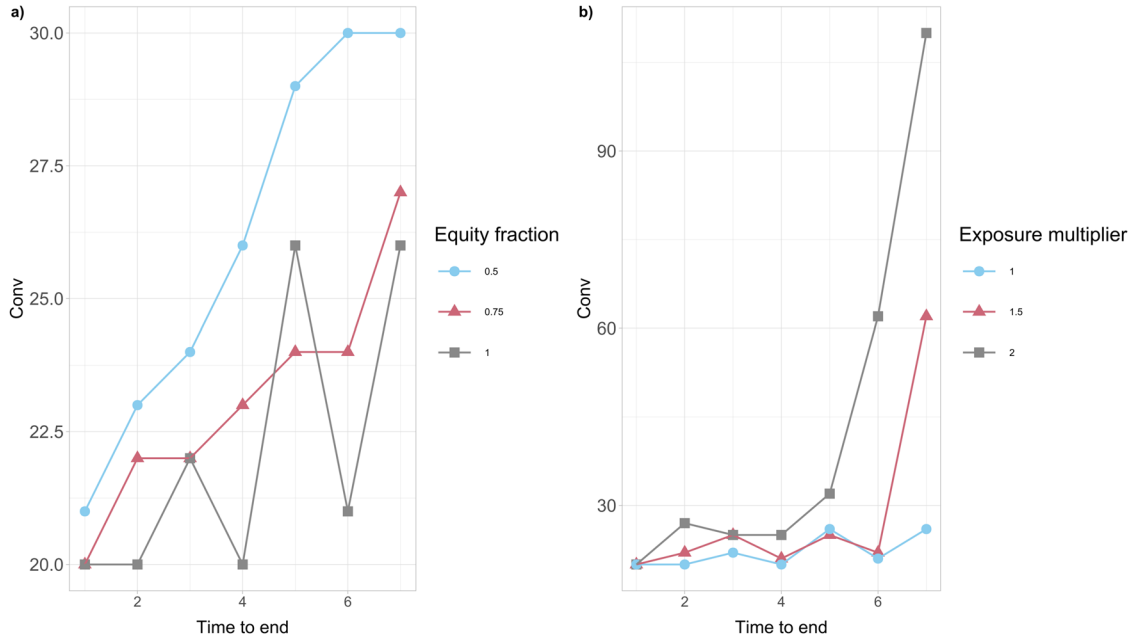

Fig. 6 | The convenience to intervene Conv for the European Banking Authority (EBA) network as a function of the time until the end of crisis. The values of Conv (in millions of EUR) are obtained for the percentage $\alpha = 0.005$ of wealth loss upon default. **a** We consider different percentages (50%, 75%, 100%) of the initial capital $E_i(0)$ available to financial institutions i at time 0. Conv tends to increase on average as the financial institutions' initial capital $E_i(0)$ decreases. It is also interesting to note that, this convenience intensifies for larger lengths of time until the end of the crisis. **b** We consider different multipliers (1, 1.5, 2) of the bilateral credit exposures w_{ij} of financial institutions i to the default of j at time 0. Conv increases as the w_{ij}'s increase and the impact of longer crisis duration on Conv is massive.

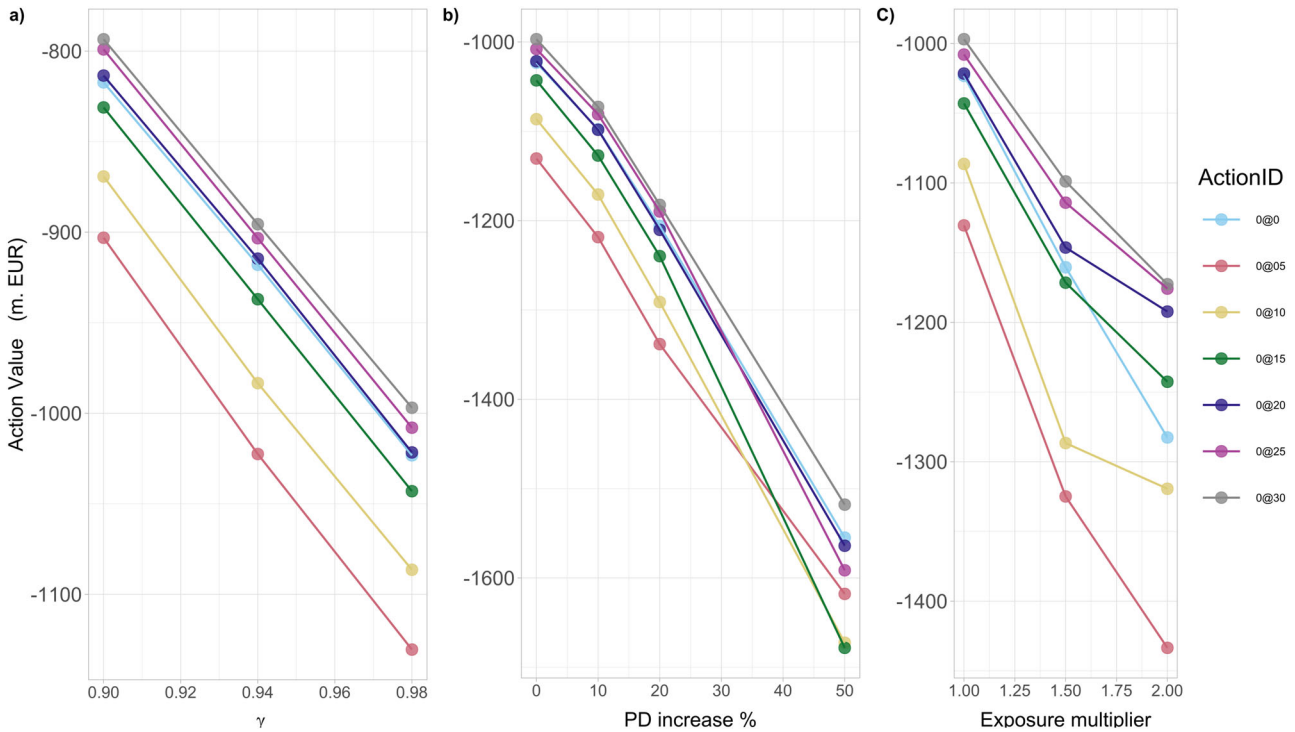

Fig. 7 | The optimal action value $Q_*(s_0, a_0)$ for the European Banking Authority (EBA) network at time 0. In the legend, ActionID 0@0 means no investment, 0@05 means investing 0.5 in all the nodes, etc. The results are obtained for the percentage $\alpha = 0.005$ of wealth loss upon default. **a** As a function of discount factor γ, $Q_*(s_0, a_0)$ decreases, for all potential actions a_0, as γ increases. **b** As a function of the percentage increase of the probabilities of default $PD_i(0)$ of financial institutions i at time 0, $Q_*(s_0, a_0)$ decreases, for all potential actions a_0, as $PD_i(0)$ increase. **c** As a function of the magnitude of multiplier of the bilateral credit exposures w_{ij} of financial institutions i to the default of j, $Q_*(s_0, a_0)$ decreases, for all potential actions a_0, as w_{ij} increase.

https://doi.org/10.1038/s41467-022-34102-1

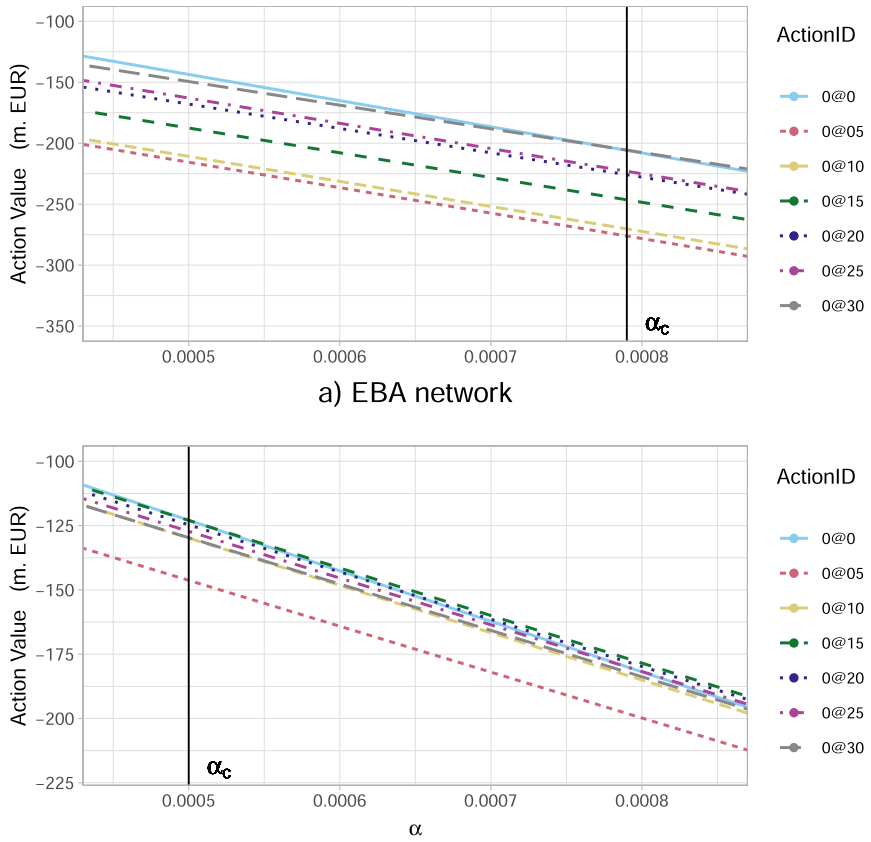

Fig. 8 | The optimal action value $Q_*(s_0, a_0)$ for the European Banking Authority (EBA) network at time 0 as a function of the percentage α of wealth loss upon default. In the legend, the ActionID 0@0 means no investment, 0@05 means investing 0.5 in all the nodes, etc. **a** There is a critical value of α given by $\alpha_c \approx 0.00079$, beyond which the inaction of the government is no longer optimal. In particular, for α just above the critical α_c, the best action becomes the investment of 3.0%W_i in all risky financial institutions i (0@30). **b** A severely distressed version of the network, where the banks' capital $E_i(0)$ has been artificially halved (all other characteristics are the same). We observe that such a distress affects $Q_*(s_0, a_0)$ at time 0 and the value α_c at which a regulatory intervention becomes favourable is lower, namely $\alpha_c \approx 0.0005$. The optimal action becomes the investment of 1.5% W_i in all risky financial institutions i (0@15).

Action value function approximation

Considering now the definition (24) of $Q_*(s_t, a_t)$ together with the approximation $\overline{V}_*(s_t,\beta)$ in (27) of $V_*(s_t)$, we introduce $\overline{Q}_*(s_t,a_t,\beta_{t+1})$ as the parametric representation of $Q_*(s_t, a_t)$, given by

$$\begin{aligned}\overline{Q}_*(s_t,a_t,\beta_{t+1}) &= \sum_{s'_{t+1}} P_{a_t}(s_t,s'_{t+1})(R_{a_t}(s_t,s'_{t+1}) + \gamma \overline{V}_*(s'_{t+1},\beta_{t+1})) \\ &= \sum_{s'_{t+1}} P_{a_t}(s_t,s'_{t+1})R_{a_t}(s_t,s'_{t+1}) + \gamma \sum_{s'_{t+1}} P_{a_t}(s_t,s'_{t+1})\overline{V}_*(s'_{t+1},\beta_{t+1}).\end{aligned} \tag{28}$$

However, the direct calculation of the above expressions (essential in the forthcoming Learning process subsection) are non-feasible in the existing form, due to the enormous set of states s'_{t+1} that can be reached from state s_t, even for relatively small networks (see the definition of the MDP for details on reachable states). In order to overcome this hurdle, we propose the following technique, which exploits the duality between the dynamics of the financial network (our nodes' default modelling) and the MDP transition probabilities and rewards. To that end, we treat the two sums on the right-hand side of eq. (28) separately, and obtain the desired reformulation in the following three steps.

Step 1. We first recall (cf. the definition (21)) that $\sum_{s'} P_a(s,s')R_a(s,s')$ is the one-step expected reward after taking action $a \in A_s$. We then recall that, the transition probability $P_a(s,s')$ was defined through the Gaussian latent variable model (see (17)) and observe that there is a one-to-one correspondence between additional nodes defaulting from state s and the resulting state s' that is reached given action a. In light of this duality, we can rewrite the first sum in terms of the nodes of the network instead of the MDP transition probabilities and rewards, according to

$$\sum_{s'_{t+1}} P_a(s_t,s'_{t+1})R_a(s_t,s'_{t+1}) = -\sum_{i\in\mathcal{I}\setminus\mathcal{I}_{def}(t)} PD_i^a(t)L_i^a(t), \tag{29}$$

with the updated probability of default $PD_i^a(t)$ and taxpayers' overall loss $L_i^a(t)$ after taking action $a_t \in A_{s_t}$ at time t, given by (13) and (16), respectively.

Step 2. The term $\sum_{s'} P_a(s,s')\ \overline{V}_*(s',\beta)$ can then be estimated via Monte Carlo simulations, which involve (a) sampling s' using the distribution P_s^a defined by the transition probability mass function $P_a(s,s')$ and (b) calculating the expected value $E^{P_s^a}[\overline{V}_*(s',\beta)]$ by averaging the values $\overline{V}_*(s',\beta)$. Hence, we have

$$\sum_{s'_{t+1}} P_{a_t}(s_t,s'_{t+1})\overline{V}_*(s'_{t+1},\beta_{t+1}) \sim E^{P_{s_t}^{a_t}}[\overline{V}_*(s'_{t+1},\beta)]. \tag{30}$$

However, the non-feasibility is essentially still present here due to the enormous number of states s'_{t+1} involved in $P_{a_t}(s_t,s'_{t+1})$, which defines the sampling distribution $P_{s_t}^{a_t}$.

Similarly to Step 1, we once again use our knowledge of the underlying network dynamics to describe the right hand side of (30)

https://doi.org/10.1038/s41467-022-34102-1

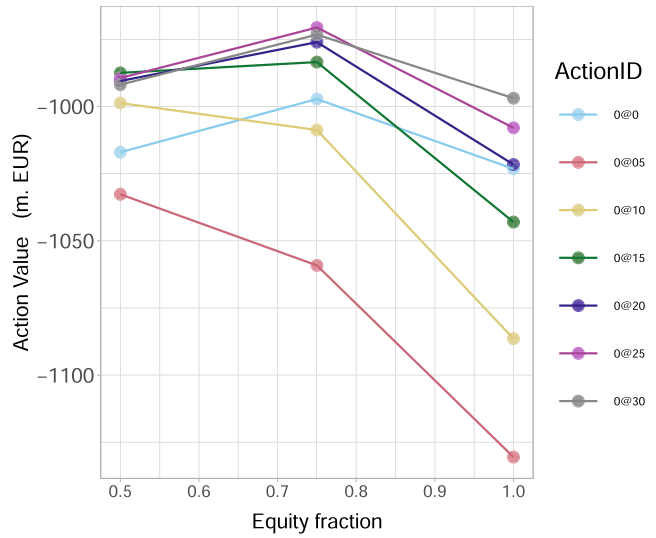

Fig. 9 | The optimal action value $Q_*(s_0, a_0)$ for the European Banking Authority (EBA) network at time 0 as a function of the percentage of initial capital $E_i(0)$ available to financial institutions i. In the legend, the ActionID 0@0 means no investment, 0@05 means investing 0.5 in all the nodes, etc. The results are obtained for the percentage $\alpha = 0.005$ of wealth loss upon default and show that the best action a^* amongst a_0's (corresponding to the highest $Q_*(s_0, a_0)$ value) decreases (0@30 → 0@25 → 0@15) as $E_i(0)$ decrease.

in terms of nodes defaulting instead of MDP transition probabilities. Using the duality between states s'_{t+1} reached given action a_t and the set $\mathcal{I}_{def}(t+1) \setminus \mathcal{I}_{def}(t)$ of additional nodes defaulting at time $t+1$, we denote by $Q^{a_t}_{s_t}$ the probability distribution of states s'_{t+1}, which are derived by using our Gaussian latent variable model. In particular, sampling from distribution $Q^{a_t}_{s_t}$ translates into first simulating which nodes i default at time $t+1$, i.e. $i \in \mathcal{I}_{def}(t+1) \setminus \mathcal{I}_{def}(t)$, via the updated default mechanism in (14), and then we obtain the corresponding state s'_{t+1} with all its resulting characteristics. Given that the distributions Q^a_s and P^a_s are equivalent (due to the aforementioned duality/one-to-one correspondence), we can therefore rewrite (30) as

$$\sum_{s'_{t+1}} P_{a_t}(s_t, s'_{t+1}) \overline{V}_*(s'_{t+1}, \beta_{t+1}) \sim E^{Q^{a_t}_{s_t}}[\overline{V}_*(s'_{t+1}, \beta)]. \tag{31}$$

Step 3. Merging the expressions obtained in Steps 1 and 2, namely (29) and (31), we can eventually rewrite $\overline{Q}_*(s_t, a_t, \beta_{t+1})$ from (28) for all $t \in [0, M-1]$, in the form of

$$\overline{Q}_*(s_t, a_t, \beta_{t+1}) = -\sum_{i \in \mathcal{I} \setminus \mathcal{I}_{def}(t)} PD^{a_t}_i L^{a_t}_i + \gamma E^{Q^{a_t}_{s_t}}\left[\overline{V}_*(s'_{t+1}, \beta_{t+1})\right]. \tag{32}$$

Learning process

In order to learn the parameters $\beta_{ik}(t)$ we primarily need to use the Bellman optimality equation from (26). Recalling the expression (25) leading to its original derivation and using the approximation $\overline{Q}_*(s_t, a_t, \beta_{t+1})$ from (28) instead of $Q_*(s_t, a_t)$, we can define a function V_B, that we call Bellman value, by

$$\begin{aligned} V_B(s_t, \beta_{t+1}) &:= \max_{a_t}\{\overline{Q}_*(s_t, a_t, \beta_{t+1})\} \\ &= \max_{a_t}\left\{\sum_{s'_{t+1}} P_{a_t}(s_t, s'_{t+1})\left(R_{a_t}(s_t, s'_{t+1}) + \gamma \overline{V}_*(s'_{t+1}, \beta_{t+1})\right)\right\}. \end{aligned}$$

or equivalently, using (32), we get the more computationally convenient (recall our proposed technique in the Action value function approximation subsection) form

$$V_B(s_t, \beta_{t+1}) := \max_{a_t}\left\{-\sum_{i \in \mathcal{I} \setminus \mathcal{I}_{def}(t)} PD^{a_t}_i L^{a_t}_i + \gamma E^{Q^{a_t}_{s_t}}\left[\overline{V}_*(s'_{t+1}, \beta_{t+1})\right]\right\}. \tag{33}$$

Our learning process will then compare our approximation $\overline{V}_*(s_t, \beta_t)$ from (27) of the optimal value function with $V_B(s_t, \beta_{t+1})$ from (33), at each state s_t and at any time t, with the aim of adjusting β so that the two values come closer. A potential issue here is that the Bellman value V_B depends itself on β, which is the parameter we want to fit, hence potentially triggering a divergent loop. In order to guarantee the convergence of our approach, we thus need to use specific learning strategies.

To begin the procedure, we can initialise β with $\beta_{ik}(t) = 1$ for all i, k, as a natural starting point due to the intuition behind our initial approximation $\overline{V}_*(s_t, \beta_t)$ of the optimal value function in (27), where $\beta_{ik}(t)$ multiply the approximated expected direct losses $\overline{Z}_{ik}(s_t)$ (recall the Value function approximation and parametrisation subsections for more details). Then we notice that, for each state s_t, our approximation $\overline{V}_*(s_t, \beta_t)$ depends on β at time t, while the corresponding $V_B(s_t, \beta_{t+1})$ on β at time $t+1$. Using this fact, in order to guarantee the convergence of our learning process, we fit β backwards in time. This results in $\overline{V}_*(s_t, \beta_t)$ being compared at time t with a value $V_B(s_t, \beta_{t+1})$ that is fixed, because β_{t+1} has already been fitted in the previous step (time $t+1$), thus solving any potential convergence problem. This learning process then repeats the same procedure backwards in time by performing a ridge regression comparing $\overline{V}_*$ with V_B, by fitting β until the difference between them is "small enough", and the procedure then concludes successfully with obtaining β^{fit}, i.e. the fitted parameters β_t for each time t.

It is worth noting that our aforementioned approach is feasible due to the facts that: (a) the crisis has a fixed maturity M, and (b) the value function $V_*(s_M) = 0$ at time M by definition (19). Consequently, we observe that $V_*(s_t) = 0$ for all $t \geq M$, while for $t \in [0, M-1]$, the backwards procedure works as follows.

The fact that $V_*(s_M) = 0$ further implies that at time $M-1$, we have $V_B(s_{M-1}, \beta_M) \equiv V_B(s_{M-1})$, since it will not depend on β. In view of (33), we can thus write

$$V_B(s_{M-1}) = \max_{a_{M-1}}\left\{-\sum_{i \in \mathcal{I} \setminus \mathcal{I}_{def}(M-1)} PD^{a_{M-1}}_i L^{a_{M-1}}_i\right\} = V_*(s_{M-1}), \tag{34}$$

where the latter equality follows from (26) and (29).

Now that we can calculate the exact optimal value function V_* for each state at time $M-1$, we notice from (33) that $V_B(s_{M-2}, \beta_{M-1}) \equiv V_B(s_{M-2})$ is also independent of β, namely

$$V_B(s_{M-2}) = \max_{a_{M-2}}\left\{-\sum_{i \in \mathcal{I} \setminus \mathcal{I}_{def}(M-2)} PD^{a_{M-2}}_i L^{a_{M-2}}_i + \gamma E^{Q^{a_{M-2}}_{s_{M-2}}}\left[V_*(s'_{M-1})\right]\right\}. \tag{35}$$

We then fit β backwards in time for the decreasing sequence of time steps $(M-2, \ldots, 0)$, by creating a representative portfolio S_R of MDP states for each time step (namely, a subset of the state space S that is reachable from s_0, see the following two subsections for details) and performing a ridge regression (with a 5-fold cross-validation) comparing $\overline{V}_*(s_t, \beta_t)$ with $V_B(s_t, \beta_{t+1})$. To be more precise, for time step $M-2$, we compare $\overline{V}_*(s_{M-2}, \beta_{M-2})$ with $V_B(s_{M-2})$, for all the states $s_{M-2} \in S_R$, and we fit β_{M-2}.

Then, for time step $M-3$, we calculate from (33) that

$$V_B(s_{M-3},\beta_{M-2}) = \max_{a_{M-3}}\left\{ -\sum_{i\in\mathcal{I}\setminus\mathcal{I}_{def}(M-3)} PD_i^{a_{M-3}} L_i^{a_{M-3}} + \gamma E^{Q^a_{s_{M-3}}}\left[\overline{V}_*(s'_{M-2},\beta_{M-2})\right]\right\}, \tag{36}$$

then compare it with $\overline{V}_*(s_{M-3},\beta_{M-3})$ and hence fit β_{M-3}, for all the states $s_{M-3}\in S_R$.

We continue the procedure backwards in time until we successfully obtain β^{fit}, i.e. the fitted β_t for each time t.

"Reachable" MDP states example

To illustrate the implementation of our model and how to identify reachable states, we consider here a simple example of a network with three nodes $\mathcal{I}=\{1,2,3\}$ and $w_{ij}=1$, for all $i\neq j\in\mathcal{I}$, at a time t. To define state s_t at time t, we assume that node $3\in\mathcal{I}_{def}(t)$ has already defaulted, while the remaining nodes have $W_i(t)=100$, $E_i(t)=3$ and $PD_i(t)=0.001$ for $i\in\mathcal{I}\setminus\mathcal{I}_{def}(t)=\{1,2\}$.

In case the government does not intervene, the states s'_{t+1} that can be reached from state s_t are those where: (i) all the nodes default at time t, i.e. $\mathcal{I}_{def}(t+1)=\mathcal{I}$; (ii) nodes 1 and 2 are still active and $W_i(t+1)$, $E_i(t+1)$ and $PD_i(t+1)$ for $i\in\{1,2\}$ are the same as for state s_t; (iii) node 1 defaults at time t while node 2 remains active, i.e. $\mathcal{I}_{def}(t+1)=\{1,3\}$, $W_2(t+1)=99$ and $E_2(t+1)=2$ (since the impact $I_2(t)=w_{21}=1$) and $PD_2(t+1)$ needs to take the value calculated via (6) using the $W_2(t+1)$ and $E_2(t+1)$ inputs; and (iv) node 2 defaults at time t but node 1 remains active, which is analogous to (iii) by swapping indices 1 and 2.

Now, if the government decides to invest, i.e. $a\to(\Delta J_1^a(t),\Delta J_2^a(t))$ on nodes 1 and 2, respectively, at time t, we need to update the capitals $E_i^a(t)=E_i(t)+\Delta J_i^a(t)$ and total assets $W_i^a(t)=W_i(t)+\Delta J_i^a(t)$ for $i\in\{1,2\}$ according to the government intervention (cf. (12)) and then use $E_i^a(t),W_i^a(t)$ to calculate the updated $PD_i^a(t)$ according to (13). Using these updated characteristics $E_i^a(t),W_i^a(t)$ and $PD_i^a(t)$, we can then perform the same analysis as above to identify the reachable states.

Representative portfolio S_R of "reachable" states

Recall that in the Value function approximation subsection, we express our approximation $\overline{V}_*(s_t,\beta)$ in (27) of the optimal value function $V_*(s_t)$ as a linear combination of terms $\overline{Z}_{ik}(s_t)$ with coefficients $\beta_{ik}(t)$. In order to fit these $\beta_{ik}(t)$, our methodology in the Learning process subsection, requires the identification of a representative portfolio S_R of MDP states that can be reached at time t from the initial state s_0, and for which we can calculate the Bellman value V_B using (33), equate it with $\overline{V}_*(s_t,\beta_t)$ and derive a set of linear equations in order to obtain the coefficients $\beta_{ik}(t)$ via a ridge regression.

The states s_t in our representative portfolio S_R at each time t, are obtained in two main ways, taking into account the trade-off between stable results and computational resources.

1^{st} way. We obtain elements $s_t\in S_R$ from the initial state s_0, after changing the time to maturity from M to $M-t$ (i.e. state s_0 is moved forward in time) and forcing a set U of nodes to default. The representative portfolio S_R contains:

(a) the state corresponding to $U=\emptyset$;
(b) all the states corresponding to $U=\{i\}$ for $i\in\mathcal{I}$, i.e. with one additional defaulted node with respect to s_0;
(c) a selection of states corresponding to $|U|>1$, i.e. with multiple additional defaulted nodes—these are chosen randomly with probabilities proportional to $exp(-|U|)$, so that a greater importance is given to states with fewer number of additional defaults, as they are more likely to be reached in an actual simulation.

2^{nd} way. In addition to the above states, we obtain elements $s_t\in S_R$ by performing a government action $a_0\in A_{s_0}$ on state s_0 and then move the corresponding state forward at time t, i.e. make the time to maturity equal to $M-t$.

Optimal solution of the MDP

Finally, we use the (fitted) optimal value function $\overline{V}_*(s,\beta^{fit})$, with β^{fit} obtained in the Learning process subsection, together with (32) to calculate $\overline{Q}_*(s,a,\beta^{fit})$, hence solving the MDP. The resulting optimal action value function is

$$Q_*(s_t,a_t)\approx\overline{Q}_*(s_t,a_t,\beta^{fit}) = -\sum_{i\in\mathcal{I}\setminus\mathcal{I}_{def}(t)} PD_i^{a_t} L_i^{a_t} + \gamma E^{Q^{a_t}_{s_t}}\left[\overline{V}_*(s'_{t+1},\beta^{fit})\right]. \tag{37}$$

Data availability

The data used in this study are available in public databases whose web links are provided in citations.

Code availability

Our research concerns a theoretical framework to solve a complex MDP that is relevant for financial systemic risk analysis. The readers can access a full mathematical description of our artificial intelligence technique and all its mathematical technicalities and their resolutions to validate our approach in the Methods section.

References

1. Haldane, A.G. Managing global finance as a system. *Bank of England, Speech at the Maxwell Fry Annual Global Finance Lecture, Birmingham University, 29*. (2014).
2. Iori, G., De Masi, G., Precup, O. V., Gabbi, G. & Caldarelli, G. A network analysis of the Italian overnight money market. *J. Econ. Dyn. Control* **32**, 259–278 (2008).
3. Eisenberg, L. & Noe, T. H. Systemic risk in financial systems. *Manag. Sci.* **47**, 236–249 (2001).
4. Glasserman, P. & Young, H. P. Contagion in financial networks. *J. Econ. Lit.* **54**, 779–831 (2016).
5. Allen, F. & Gale, D. Systemic risk and regulation. *The Risks of Financial Institutions*, 341–376 (University of Chicago Press, 2007).
6. Natwest Group. Equity ownership statistics of the Natwest Group. Available at https://investors.natwestgroup.com/share-data/equity-ownership-statistics.aspx (2020).
7. Office for Budget Responsibility. Economic and fiscal outlook. Available at https://obr.uk/download/economic-and-fiscal-outlook-march-2018/ (2018).
8. National Audit Office. Maintaining financial stability across the UK's banking system. Available at https://www.nao.org.uk/wp-content/uploads/2009/12/091091.pdf (2009).
9. Furfine, C. Interbank exposures: Quantifying the risk of contagion. *J. Money Credit Bank.* **35**, 111–128 (2003).
10. Lehar, A. Measuring systemic risk: A risk management approach. *J. Bank. Finance* **29**, 2577–2603 (2005).
11. Davis, M.A. & Lo, V. Modelling default correlation in bond portfolios. In *Mastering Risk, Volume 2: Applications*, 141–151 (2001).
12. Caccioli, F., Barucca, P. & Kobayashi, T. Network models of financial systemic risk: a review. *J. Comput. Soc. Sc.* **1**, 81–114 (2018).
13. Gai, P. & Kapadia, S. Contagion in financial networks. *Proc. Royal Soc. A* **466**, 2401–2423 (2010).
14. Battiston, S., Puliga, M., Kaushik, R., Tasca, P. & Caldarelli, G. DebtRank: too central to fail? Financial networks, the FED and systemic risk. *Sci. Rep.* **2**, 541 (2012).
15. Strogatz, S. H. Exploring complex networks. *Nature* **410**, 268–276 (2001).
16. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M. & Hwang, D. Complex networks: Structure and dynamics. *Phys. Rep.* **424**, 175–308 (2006).

17. Leduc, M. V. & Thurner, S. Incentivizing resilience in financial networks. *J. Econ. Dyn. Control* **82**, 44–66 (2017).
18. Bardoscia, M., Battiston, S., Caccioli, F. & Caldarelli, G. Pathways towards instability in financial networks. *Nat. Commun.* **8**, 14416 (2017).
19. Gerhardt, M. & Vander Vennet, R. Bank bailouts in Europe and bank performance. *Finance Res.* **22**, 74–80 (2017).
20. Carbó-Valverde, S., Cuadros-Solas, P. J. & Rodríguez-Fernández, F. Do bank bailouts have an impact on the underwriting business? *J. Financial Stab.* **49**, 100756 (2020).
21. Lin, K. L., Wu, T. C. & Li, K. P. Government support of banks and market discipline: International evidence. *Rev. Financial Econ.* **40**, 174–199 (2022).
22. Cuadros-Solas, P. J., Salvador, C. & Suárez, N. Am I riskier if I rescue my banks? Beyond the effects of bailouts. *J. Financial Stab.* **56**, 100935 (2021).
23. Laeven, L. & Levine, R. Bank governance, regulation and risk taking. *J. Financ. Econ.* **93**, 259–275 (2009).
24. Brandao-Marques, L., Correa, R. & Sapriza, H. Government support, regulation, and risk taking in the banking sector. *J. Bank. Financ.* **112**, 105284 (2020).
25. Nistor, S. & Ongena, S. The impact of policy interventions on systemic risk across banks. *Swiss Finance Institute Research Paper* (20–101) (2020).
26. Berger, A. N., Nistor, S., Ongena, S. & Tsyplakov, S. Catch, restrict, and release: The real story of bank bailouts. *Swiss Finance Institute Research Paper* 20–45 (2020).
27. Veronesi, P. & Zingales, L. Paulson's gift. *J. Financ. Econ.* **97**, 339–368 (2010).
28. Minca, A. & Sulem, A. Optimal control of interbank contagion under complete information. *Stat. Risk Model.* **31**, 23–48 (2014).
29. Amini, H., Minca, A. & Sulem, A. Control of interbank contagion under partial information. *SIAM J. Financ. Math.* **6**, 1195–1219 (2015).
30. Amini, H., Minca, A. & Sulem, A. Optimal equity infusions in interbank networks. *J. Financ. Stab.* **31**, 1–17 (2017).
31. Demange, G. Contagion in financial networks: a threat index. *Manag. Sci.* **64**, 955–970 (2018).
32. Capponi, A., Corell, F. C. & Stiglitz, J. E. Optimal bailouts and the doom loop with a financial network. *J. Monet. Econ.* **128**, 35–50 (2022).
33. Upper, C. & Worms, A. Estimating bilateral exposures in the German interbank market: Is there a danger of contagion? *Eur. Econ. Rev.* **48**, 827–849 (2004).
34. Upper, C. Simulation methods to assess the danger of contagion in interbank markets. *J. Financ. Stab.* **7**, 111–125 (2011).
35. Cont, R. & Wagalath, L. Running for the exit: distressed selling and endogenous correlation in financial markets. *Math. Finance* **23**, 718–741 (2013).
36. Souza, S. R. Sd, Silva, T. C., Tabak, B. M. & Guerra, S. M. Evaluating systemic risk using bank default probabilities in financial networks. *J. Econom. Dynam. Control* **66**, 54–75 (2016).
37. Petrone, D. & Latora, V. A dynamic approach merging network theory and credit risk techniques to assess systemic risk in financial networks. *Sci. Rep.* **8**, 5561 (2018).
38. Merton, R. C. On the pricing of corporate debt: The risk structure of interest rates. *J. Finance* **29**, 449–470 (1974).
39. Bellman, R. E. A Markovian decision process. *J. Math. Mech.* **6**, 679–684 (1957).
40. Gordon, G. Approximate solutions to Markov Decision Processes. PhD thesis, Carnegie Mellon University Pittsburgh, PA. (1999).
41. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction* (MIT Press, 2018).
42. Anand, K., Craig, B. & Von Peter, G. Filling in the blanks: Network structure and interbank contagion. *Quant. Finance* **15**, 625–636 (2015).
43. Black, F. & Cox, J. C. Valuing corporate securities: Some effects of bond indenture provision. *J. Finance* **31**, 351–367 (1976).
44. O'Kane, D. The gaussian latent variable model in *Modelling Single-name and Multi-name Credit Derivatives* (Wiley Finance, 2008) 241–259.
45. Flanagan, T. & Purnanandam, A. Did Banks Pay 'Fair' Returns to Taxpayers on TARP? *Available at SSRN 3595763* (2020).
46. Bellman, R. E. *Dynamic Programming*. (Princeton University Press, 1957).
47. Heynderickx, W., Cariboni, J., Schoutens, W. & Smits, B.F. European Banks' Implied Recovery Rates. *Available at SSRN 3595763* (2016).
48. Huang, X., Zhou, H. & Zhu, H. A framework for assessing the systemic risk of major financial institutions. *J. Bank. Finance* **33**, 2036–2049 (2009).
49. Credit Suisse. Annual report 2019, Credit Suisse counterparty ratings (p.147). Available at https://www.credit-suisse.com/media/assets/corporate/docs/about-us/investor-relations/financial-disclosures/financial-reports/csg-ar-2019-en.pdf (2020).
50. Krackhardt, D. Assessing the political landscape: Structure, cognition, and power in organizations. *Adm. Sci. Q.* **35**, 342–369 (1990).
51. Arya, A. & Glover, J. Excessive intervention, the collusion problem, and information system design (2001).
52. Acharya, V. V. & Yorulmazer, T. Too many to fail? An analysis of time inconsistency in bank closure policies. *J. Financ. Intermed.* **16**, 1–31 (2007).
53. Dam, L. & Koetter, M. Bank bailouts and moral hazard: Evidence from Germany. *Rev. Financ. Stud.* **25**, 2343–2380 (2012).
54. Calomiris, C. W. & Jaremski, M. Stealing deposits: Deposit insurance, risk-taking and the removal of market discipline in early 20th-century banks. *J. Finance* **74**, 711–754 (2019).
55. European Banking Authority. Global Systemically Important Institutions, 2014 data. Available at https://eba.europa.eu/risk-analysis-and-data/global-systemically-important-institutions (2015).
56. Fitch credit rating agency. Available at https://www.fitchratings.com (2021)

Author contributions

D.P. devised and performed the research, conceived the model and wrote the paper; N.R. devised the research, mathematically defined and structured the model and wrote the paper; V.L. devised the research, reviewed the paper and improved its structure and content.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Neofytos Rodosthenous.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.