Citizen preferences and the architecture of government

Jean-Marc Bourgeon, Marie-Laure Breuillé

Citizen preferences and the architecture of government

Jean-Marc Bourgeon* Marie-Laure Breuillé†

July 8, 2021

Abstract

We consider the division of a territory into administrative jurisdictions responsible for providing a set of goods to residents who are sensitive to congestion effects. We deduce the optimal architecture of public governance (i.e. the division of government into several levels, the distribution of services among them, their number of jurisdictions and the capacity of their administrations), which depends on citizens preferences regarding the quality of public services. We compare it to a decentralized organization where each jurisdiction is free to choose the capacity and scope of its administration. The resulting architecture of government generally involves more countries with fewer levels of administration than the optimal one. Our results allow us to estimate citizen preferences for the U.S. We find that the country is divided into two zones ("Northeast & West" and "Midwest & South") whose estimated values are statistically different.

Keywords: Decentralization, Fiscal Federalism, Public Governance.

JEL classification: H11, H77

*INRAE, UMR économie Publique, 16 rue Claude Bernard, F-75231 Paris, France, and CREST - Ecole Polytechnique, France. Email: bourgeon@agroparistech.fr.

†INRAE, UMR1041 CESAER, Université Bourgogne Franche-Comté, AgroSup Dijon, F-21000 Dijon, France. Email: marie.breuille@inrae.fr.

We are grateful to Julie Le Gallo, Anne-Laure Samson, to seminar participants at the economic department of Ecole Polytechnique, TSE, BETA, the Goethe University Frankfurt, the UCLouvain Belgo-Japanese Public Finance workshop and the CESIfo Applied Microeconomics and Political Economy workshops.

1 Introduction

The architecture of government (AG), i.e. the organizational layout and functionality of a country's public administration in totality (the number of its tiers and jurisdictions per tier, their geographical distribution and the services they are in charge of providing), differs widely over the world. The number of tiers of sub-national authorities ranges from five, such as in the Philippines, to only one, such as in Kuwait. The number of jurisdictions in each tier also varies widely: focusing on the bottom-most tier of municipal authorities in Europe, 40% are located in a single country: France. The average size of municipalities varies from less than 2,000 inhabitants, as in Mongolia and the Czech Republic, to 130,000 inhabitants for African and Asia-Pacific countries (OECD/UCLG, 2019). Most countries have planned or completed reforms (often guided by cost reduction considerations and socio-demographic changes) to redraw the map of AGs in the past 30 years, through municipal-mergers, inter-municipal cooperation, or metropolitan governance. Denmark totally reorganized its AG in 2007; the number of municipalities dropped from 271 to 98 through a series of mergers, and its 13 counties were replaced by 5 regions (Dexia, 2007). Latvia reduced the number of municipalities from 527 to 105 in 2009, and France its number of metropolitan regions from 22 to 13 in 2016. In parallel, OECD (2019) points the increasing power of regions, with 52 countries out of 81 experiencing a significant increase in the degree of regional authority since the 1970s. Some developing countries, such as Bolivia in 1995 and Indonesia in 1998, have also undertaken major decentralization reforms (Bardhan & Mookherjee, 2006). The diversity of forms of public governance that we observe is admittedly the outcome of country-specific differences in geography, history, and political and social movements, but for "making decentralization work" (OECD, 2019), the question of the desirable properties of an AG remains open.

In this paper, we examine this issue both theoretically and empirically by focusing on citizens' expectations regarding the quality of public services. Indeed, depending on the number of tasks that an administration has to perform and the means at its disposal, more or less significant congestion effects affect its performance (leading to delays, queues, inadequacies or breakdowns in the provision of services) that can frustrate the public. We first develop a theoretical framework that takes into account citizens' expectations on these qualitative aspects of public activity to derive the optimal AG, i.e. for a given territory, the number of countries, tiers and jurisdictions per tier, the capacity of the administrations, and the goods they produce at each level. We derive the AG that would be designed and administrated by a benevolent social planner aiming at maximizing the social welfare and compare this socially optimal AG to a "decentralized" form of government where sub-national jurisdictions have some leeway in determining the scope and the capacity of their administration. We then use the expressions derived from this theoretical framework to perform an empirical assessment of citizens' preferences using U.S. data that shows that the country is divided into two zones ("Northeast & West" and "Midwest & South") where residents' expectations regarding the quality of public services are statistically different.

Our theoretical setup consists in a world modeled as a linear segment populated by a continuum of residents uniformly distributed and identified by their geographical locations. Preferences of individuals depend on the bundle of goods produced and on the productive capacity of the public facilities, in addition to their localization and production costs, to reflect the trade-off between the sheer number of goods and the service quality an administration offers to its users.[1] A government can apportion the public task between different administrations serving a more of less large share of the population, leading to different levels of administrative jurisdictions. We allow for as many countries and levels of sub-national jurisdictions as necessary so that each government can best serve its citizens according to their expectations. This leads to multi-tier AGs where the linear segment is partitioned at each layer into jurisdictions, each of them in charge of producing the tier-specific bundle of public goods. Each public good is characterized by a parameter that determines the citizens' access cost, which is a function of the distance between a resident and the location of the corresponding public facility, and its production cost. The production cost of an administration depends both on the jurisdiction's scope (i.e., the set of public goods produced) and its productive capacity (i.e., the capital and labor involved). Hence, increasing the performance ability of the administration comes with higher production costs. Economies of scope can be achieved by increasing the number of goods produced by a jurisdiction, but at the expense of a lower quality of service. The optimal balance depends on citizen preferences and affects the overall AG.

The architecture of governments also depends on the leeway given to jurisdictions. We consider two cases: the first-best situation where all jurisdictions are ruled by a benevolent and omniscient planner, and a decentralized organization where decision-makers of each jurisdiction level decide on the scope of activities and the capacity of their administration to maximize the welfare of their constituency without accounting for the decisions made by other tiers. We derive the optimal AG in the latter case by

[1]In the following, we use interchangeably "productive capacity" and "performance ability" of the administration to refer to its capacity to contain congestion effects.

assuming that a social planner decides on the numbers of countries, tiers and jurisdictions per tier, and makes sure that there is no overlap of activities, i.e., no public good is produced by jurisdictions belonging to two different tiers of the same country. It is a second-best optimum because the social planner delegates to jurisdictional authorities with narrow objectives the management of their administration. We assume in both cases that the budget of each jurisdiction is balanced and that its production cost is financed by a lump-sum tax.

The resulting AGs are driven by four parameters, one related to the valuation of the public activity (the weights citizens assign to the administration's capacity and scope of activities are given by β and $1-\beta$ respectively), one that reflects the preference for a simple AG, and the last two related to the production cost of the administration and the residents' access cost. We first derive the characteristics of the one-tier equivalent of the nations formation model of Alesina & Spolaore (1997, hereafter AS). While an increase of the production cost diminishes the number of countries in AS, it decreases the administration's capacity of each country but increases their number in our one-tier equivalent. Hence, the capacity of each administration is reduced, but the resulting increase in congestion effects is somehow compensated for by more countries allowing a reduction in the cost of citizens' access to their government. Considering the effect of a change in the citizens' preferences, the number of countries decreases when β increases because of the increase in the production cost due to the increase in the administrations' capacity.

Allowing for an apportionment of the public tasks between several administrative tiers, the shape of the first-best AG depends on β being larger or smaller than $1/4$.[2] When $\beta = 1/4$, the per capita costs of access and production of the administrations as well as the citizen's evaluation of their activity are the same for all tier (except the last one). Proceeding down the AG, the administration's scopes increase, their capacities decrease (both in absolute and per capita terms) and the degree of jurisdictional dispersion (i.e. the ratio of the numbers of jurisdictions of two consecutive tiers) increases. Static comparative exercises around $\beta = 1/4$ show that compared to that case, when $\beta > 1/4$ the number of countries and jurisdictions per level increases, the perimeter of the central government is reduced while those of the other tiers are increasingly wider and the satisfaction that citizens derive from the activity of each level of government decreases. This is the reverse when $\beta < 1/4$, with the scope of the central government being enlarged compared to $\beta = 1/4$, sub-national tiers having decreasing scopes

[2]This threshold corresponds to the mid-value of the relevant range for β in our model, $\beta > 1/2$ resulting in access and production costs larger than the value citizens place on public activity.

and citizen welfares increasing. In summary, all AGs have a large administration at the central government level and decreasing administration sizes at sub-national tiers proceeding downward, but when individuals are not overly sensitive to the quality of public services ($\beta < 1/4$), the number of countries is small, the scope of their central government is large, and the number of sub-national tier jurisdictions is small. On the contrary, when individuals expect good public services quality ($\beta > 1/4$), the numbers of countries and jurisdictions per level are large and the central government's scope is small.

Allowing for decentralized decision-making, i.e. jurisdictions that are free to select their range of services and the capacity of their administration to perform them, we obtain that the jurisdiction scopes relative to their administration's capacity is larger than under the first-best AG. Indeed, while the social planner arbitrates between two consecutive tiers to determine their scopes, jurisdictions' choices under decentralization are based solely on the satisfaction of their residents with their own activity, neglecting the detrimental impact of increasing their scope on the lower tier in the AG. To limit this negative effect, the social planner modifies the AG by adjusting the number of tiers and jurisdictions per tier. Hence, decentralization affects the whole AG, which generally involves more states with fewer levels of administration than the optimal one.

The results of our structural model permit estimation of the citizen preference parameter β using standard panel data procedures. We conducted this empirical investigation using U.S. data on annual federal and state expenditures, share of population, and density (together with data on income per capita and on the partisan composition of the state legislatures) over the period 1977-2015. Assuming the same citizen preferences across the country, we obtain an overall estimate for β equal to 0.18, a value that is neither influenced by the partisan composition of state legislatures nor by other characteristics such as income and population density. However, disaggregating the country according to the four regional divisions used by the Census Bureau (Midwest, Northeast, South, West), we obtain that this weight is lower in the Northeast and West (0.182 and 0.147) than in the Midwest and South (0.279 and 0.238). Gathering these four regions into two zones, we obtain 0.153 for the Northeast & West zone and 0.251 for the Midwest & South zone, estimates that are statistically different. We can therefore infer that citizens in the Northeast and West regions are more appreciative of state-provided services than those provided by the federal government, whereas citizens in the Midwest and South are more likely to be equally satisfied with both levels of government.

Our paper builds a bridge between the literature on the formation of jurisdictions and the literature on fiscal federalism by formalizing the vertical dimension of the formation of jurisdictions. The literature on the formation of jurisdictions, which analyzes the equilibrium partition of a population into several jurisdictions and their political stability, has been quite extensive over the last twenty years. The breakup or unification of nations has been modeled as the result of a trade-off either between the efficiency gains of unification and the costs in terms of loss of control in political decision making (Bolton & Roland, 1997) or, more commonly, between benefits from economies of scale in the production of public goods and costs from preferences heterogeneity (Alesina & Spolaore (1997), Jehiel & Scotchmer, 2001). A cooperative game theory approach is usually used to study the political stability of jurisdictions (Guesnerie & Oddou, 1981; Greenberg & Weber, 1986; Demange, 1994; Casella, 2001; Bogomolnaia et al., 2006, 2008, among others). Although the coexistence of several clubs of different sizes was studied by Hochman et al. (1995), the vertical dimension of the formation of jurisdictions, i.e. the partition of the country into overlapping tiers, is missing in this literature. By contrast, some recent works in the fiscal federalism literature focus on the vertical dimension to determine the optimal level of decentralization of a unique public good (Panizza, 1999) or the partition of a continuum of identical public goods (Wilson & Janeba (2005), Janeba & Wilson, 2011). Their framework, however, is restricted to two tiers, i.e., a central government and an exogenous number of same-tier sub-national jurisdictions, whereas we allow for an endogenous number of tiers among which the bundle of heterogenous public goods is shared to compare the decentralized AG to the first-best one. Hierarchies in firms have been widely studied, whether as a means of information processing (Radner, 1992; Van Zandt, 1999) resource allocation (Geanakoplos & Milgrom, 1991; Van Zandt, 2003), incentive-based worker control (Calvo & Wellisz, 1978; Qian, 1994; Chen, 2017), or the acquisition of expertise to solve production problems (Garicano, 2000; Caliendo & Rossi-Hansberg, 2012). While these articles also endogenously determine the number of levels and entities (workers) per level, the forces shaping the structure of firms in these studies differ greatly from those shaping the architecture of governments, in part because of the spatial dimension of the problem. Each local jurisdiction is a sub-segment of the territory, and the location of its production site affects the cost of access for citizens.

The rest of the paper is organized as follows. Section 2 presents the multi-tier framework. Section 3 characterizes the optimal AG. Section 4 investigates the impact of decentralized decision-making. The empirical investigation is detailed in section 5. The last section concludes. All proofs are in the appendix.

2 The model

Following the spatial approach of Hotelling (1929), consider a territory with a continuum of residents uniformly distributed over a segment. The size of the territory and the population mass are both normalized to unity. Each resident is identified by her/his geographical location, i.e. a point on the segment, supposed to be fixed. We are interested in both the horizontal and vertical organization of governments, i.e. the numbers of jurisdictional tiers and jurisdictions per tier, that are responsible for providing a set of public goods to these residents. Some of these goods are readily accessible by all, such as the protection provided by the army of the country and the foreign affairs and intelligence services, while others, such as the natural amenities found in a park or the education dispensed in an elementary school, are enjoyable more or less depending on their locations relative to the residents. To capture these heterogeneity and accessibility issues in a tractable way, we consider that the public task is to provide a continuum of goods $[1, \overline{x}]$, parameter $x \in [1, \overline{x}]$ characterizing the accessibility and production costs of the corresponding public good, both of which being larger the larger this parameter.[3]

First consider the access cost. Public goods produced by an administration are available to each citizen on the site of the corresponding public facility. The cost borne by a resident to access to good $x \in [1, \overline{x}]$ produced by a public facility located at distance ℓ from her home being given by $\alpha x \ell$ where $\alpha > 0$ is the access cost parameter.[4] To handle this accessibility problem, a government is composed of jurisdictional tiers, indexed by $t \in \{0, ..., T\}$, charged to produce a subset of these services. Tier $t = 0$ is the top level, the central/federal jurisdiction, which produces the most accessible public goods, while $t = T$ is the bottom-most one, which may correspond to villages or the districts of big cities, in charge of producing the less accessible goods. Each citizen thus belongs to $T + 1$ overlapping jurisdictions, i.e. T sub-national jurisdictions plus the central jurisdiction, each being responsible for delivering a specific bundle of services. The tier t's task is to produce public goods belonging to the subset $(x_{t-1}, x_t]$,

[3] The extent of the rivalry and excludability characteristics of theses goods/services may vary and some of them may even be private goods. We do not discuss in the following the overall scope of the services $[1, \bar{x}]$ and we suppose that the exogenous upper bound $\bar{x}$ is the cut-off between the public and the private sectors.

[4] In the examples given above, military protection corresponds to an access cost close to 0 (arguably strictly positive for citizens living on the borders of the territory) and thus we may expect $0 < \alpha < 1$. Observe that at equal distance from home, the accessibility parameter of children's education may be larger than the one associated to natural amenities offered by a park due to inherent constraints, such as the frequency of trips to and from the school and the ease with which parents can conform to the school's operating hours.

with $x_{-1} = 1$ and $x_T = \overline{x}$. More precisely, at each level t, the territory is divided in n_t jurisdictions indexed by z_t, $z_t \in \{1, \ldots, n_t\}$, the case $n_0 \geq 2$ corresponding to a territory divided in two countries or more, each having its own AG. Jurisdiction z_t rules over the area $S_{z_t} \subseteq [0,1]$, with $S_{z_t} \cap S_{z'_t} = \varnothing$ and $\cup_{z_t=1}^{n_t} S_{z_t} = [0,1]$, and is geographically defined by three points: its two borders and the location ρ_{z_t} of its public facility which is also where its residents have to go to enjoy the public goods produced by its administration. Hence, to obtain the whole set of public goods, citizen i incurs a total access cost $\alpha \sum_{t=0}^{T} \ell_{it} \int_{x_{t-1}}^{x_t} x dx$, where $\ell_{it} \equiv |i - \rho_{z_t}|$ represents the distance to reach the location ρ_{z_t} of the public facility in tier t. Figure 1 gives an example of AG.

Next, consider the production cost of the government. Each jurisdiction entails a production cost that depends on the range of the public goods it provides and the performance ability of its administration. We suppose that this performance depends on the resources used for the production (staff, size of the public facility, amount of capital), i.e. the administration's capacity.[5] As noted above, we assume that the production cost of a good is larger the higher its accessibility parameter x regardless of the capacity of the administration.[6] However, an administration can achieve economies of scope by producing a wide range of public goods. More precisely, denoting by $c(g, x)$ the cost of producing good x by an administration of capacity g, the total production cost of the bundle $(x_{t-1}, x_t]$ of services by the same administration of capacity g_{z_t} is given by

$$C_{z_t} = C(g_{z_t}, (x_{t-1}, x_t]) = \int_{x_{t-1}}^{x_t} c(g_{z_t}, x) dx \Big/ \int_{x_{t-1}}^{x_t} dx.$$

Without economies of scope, the production cost would be given by $\int_{x_{t-1}}^{x_t} c(g_{z_t}, x) dx$. Economies of scope arise from having the same administration to produce the full range of services $\int_{x_{t-1}}^{x_t} dx = x_t - x_{t-1}$. We suppose that $c(g, x)$ increases with the accessibility parameter and the capacity of the administration: $c'_x(g, x) > 0$ and $c'_g(g, x) > 0$.[7] To ease computations, we consider the case $c(g, x) = kgx$, with $k > 0$, which yields $C_{z_t} = kg_{z_t}(x_t + x_{t-1})/2$. In this expression, the production cost is proportional to the

[5]In addition to the construction or the rental cost of the public facility, large administrations receiving the public necessitate more staff (administrative, security and maintenance staff) and equipment than small ones.

[6]For instance, the per capita cost of country's military or intelligence services is lower than the per capita cost of primary school teachers.

[7]Cost function $c(\cdot)$ gives the cost of producting a good potentially available to all citizens. Because of access costs, an individual opts to patronize the nearest facility that produces it, and thus, depending on the fraction of the population that attends a given facility, the per-capita cost of public services varies. Since $c(\cdot)$ increases with x, as does the access cost (which forces the government to replicate its production across the territory), the larger x is, the greater the per capita cost of providing it.

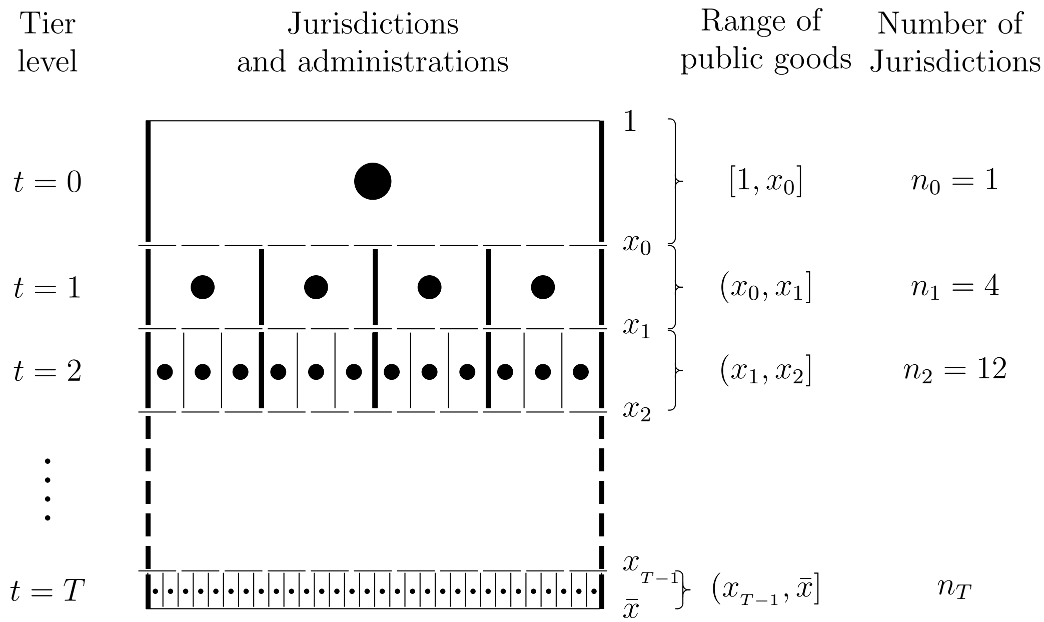

Figure 1: Example of government architecture. The height of the box corresponds to the total range of public goods: $[1, \bar{x}]$. Its length is the size of the territory (normalized to 1) that belongs to a single country ($n_0 = 1$). Horizontal segments delineate tiers, with vertical distances between them corresponding to the range of public goods produced at each level. The dots represent the location and the capacity of public facilities. The first tier, $t = 0$, is the central government that provides a range $[1, x_0]$ of goods for the entire country. Its administration has a large capacity and is centrally located. The second level, $t = 1$, is composed of 4 regions of same size, their public facilities are centrally located and produce public goods $(x_0, x_1]$ for their residents. Their capacity is smaller than that of the central government. The third tier, $t = 2$, entails 12 departments, 3 for each region. These subdivisions continue down to the last tier level T (e.g. the city level) that produces the remaining range of public services, $(x_{T-1}, \bar{x}]$.

product of the administration's capacity and the average accessibility parameter of the bundle of goods produced. Total production costs are covered by a lump-sum tax τ_{z_t} paid by each resident of the jurisdiction z_t. As tier-t encompasses n_t jurisdictions, the aggregated cost induced by the production of all public goods nationwide amounts to $\sum_{t=0}^{T}\sum_{z_t=1}^{n_t} C_{z_t}$.

Each citizen derives a utility from both the public goods provided and the administration's capacity of each of the $T+1$ jurisdictions to which he or she belongs. Indeed, in order to consume public goods, the citizens of a jurisdiction must be at the public facility that produces them, and are therefore subject to congestion effects that can be reduced by increasing the administration's capacity. For simplicity, assume that the utility of individuals at the tier-t public facility increases with the range of goods it produces,[8] and that congestion costs are proportional to this benefit. More precisely, denoting by θ_{z_t} the share of welfare lost at facility z_t due to these detrimental effects, the utility (expressed in monetary terms) from the services $(x_{t-1}, x_t]$ provided by this tier-t administration net of congestion effects is given by

$$u_{z_t} = (1-\theta_{z_t})(x_t - x_{t-1}). \tag{1}$$

Congestion effects increase with the range of services produced by the administration, due e.g. to time individuals have to spend on site to be served, but they are reduced by an increase in the administration's capacity, g_{z_t}, i.e. the larger the administration's capacity, the easier it is for citizens to enjoy the goods and services it produces.[9] The resulting share of losses is given by

$$\theta_{z_t} = 1 - \left(\frac{g_{z_t}}{x_t - x_{t-1}}\right)^{\beta}, \tag{2}$$

where $0 < \beta < 1$, with relevant values $g_{z_t} \leq x_t - x_{t-1}$ so that θ_{z_t} belongs between 0 and 1. Hence, if the administration capacity is low, i.e. g_{z_t} close to 0, the congestion cost is so large that citizens derive almost no benefit from public services. Observe that β is not a technical parameter but rather captures the citizens' appreciation of the goverment's effort to reduce congestion effects. This is apparent when we replace

[8] We neglect possible complementarity/substitutability between public goods, so that only the sheer number of goods is relevant.

[9] For example, with more staff in the public facility, it can be expected that queues will be reduced and paperwork completed in a timely fashion. Or, with more spending on public transport, a city may increase the frequency of its buses.

(2) in (1) which gives

$$u_{z_t} = \left(\frac{g_{z_t}}{x_t - x_{t-1}}\right)^{\beta}(x_t - x_{t-1}) = g_{z_t}^{\beta}(x_t - x_{t-1})^{1-\beta}.$$

In this expression of the monetary evaluation of the activity of the z_t jurisdiction, β and $1-\beta$ reflect the importance that individuals place on the administration's capacity and its scope of services, respectively, due to the impact of congestion effects.[10] Without congestion effects ($\beta = 0$), the citizen's total utility $\sum_{t=0}^{T} u_{z_t}$ would depend only on the total range of public services, $\bar{x} - 1$, whatever the AG. The capacity of the administration would be minimal and only the access cost would matter. Accounting for the congestion cost ($\beta > 0$) introduces a trade-off between the scope of each jurisdiction and the capacity of its administration due to their impact on the provision cost.[11] Economies of scope reduce the spending related to the range of services to a function of the average accessibility parameter $(x_t + x_{t-1})/2$ of the bundle of good produced by a facility, which would be equal to $(\bar{x} - 1)/2$ if the government were reduced to a single administration ($T = 0$). However, since the cost of accessing services differs across goods, the government is compelled to divide activities among increasingly local jurisdictions. As the average accessibility parameter increases the more local the administration, the capacity of their facility is reduced to contain the provision cost. Increasing the number of tiers could therefore allow the government to reduced the access cost while minimizing the provision cost. We suppose however that such a process is limited by a setup cost of the corresponding AG that increases with the number of tiers added to the central government, given by νT.[12]

We derive in the following the AG resulting from this trade-off assuming a decentralized process that we compare to an optimal organization. The first-best is achieved assuming that this arbitrage is resolved by a benevolent social planner who considers the whole AG to determine the jurisdictions size, scope and administration's capacity at each tier level. In the context of decentralization, this arbitrage is resolved sequentially, from the highest to the lowest level, with each jurisdiction being allowed to choose the capacity of its administration and its scope among the tasks not yet chosen.

[10] Alternatively, g_{z_t} could be considered as a measure of the overall quality of the administration, which depends on both the public infrastructure and the staff employed.

[11] As total utility must be larger than the production and access costs, congestion costs cannot be too large, implying $\beta \leq 1/2$ in our setting (see the appendix or footnote 14 for more details).

[12] Alternatively, we may consider that citizens have a preference for a simple AG, i.e. one that involves a reduced number of levels, since the more complex the government structure, the more difficult it is to know which administration to turn to for a given service.

3 The socially optimal AG

Consider a benevolent social planner aiming at maximizing the social welfare, who chooses the desirable division of the territory into countries and sub-national tiers, the geographical boundaries of each jurisdiction, the location and the capacity of their administration, and the range of public goods they produce. The corresponding optimization problem is given by

$$\max_{T,\{x_t,n_t,\{S_{z_t},\rho_{z_t},g_{z_t}\}_{z_t=1}^{n_t}\}_{t=0}^{T}} \sum_{t=0}^{T}\sum_{z_t=1}^{n_t}\left[\int_{S_{z_t}}\left(g_{z_t}^{\beta}\left(x_t-x_{t-1}\right)^{1-\beta}-\alpha\ell_{it}\int_{x_{t-1}}^{x_t}xdx\right)di-C_{z_t}\right]-\nu T \tag{3}$$

under $x_{-1}=1$, $x_T=\overline{x}$ and $\sum_{z_t=1}^{n_t}S_{z_t}=1$ for all t. Because individuals are uniformly located in the territory, it is easily shown that

Lemma 1 *At each tier t, the territory is divided into n_t jurisdictions of equal size. Their administrations are located at the center of the jurisdiction and have the same capacity.*

Due to the symmetry of the jurisdictions at each tier level, they have the same production cost and administration's capacity, i.e. $C_{z_t}=C_t$ and $g_{z_t}=g_t$ for all t, and the average access cost of their residents is the same. The social planner's program can be restated as

$$\max_{T,\{n_t,x_t,g_t\}_{t=0}^{T}}\left\{\sum_{t=0}^{T}W(x_{t-1},x_t,g_t,n_t)-\nu T : x_{-1}=1, x_T=\overline{x}\right\} \tag{4}$$

where νT corresponds to the setup cost of the AG and

$$W(x_{t-1},x_t,g_t,n_t)\equiv g_t^{\beta}\left(x_t-x_{t-1}\right)^{1-\beta}-\frac{\alpha\left(x_t^2-x_{t-1}^2\right)}{8n_t}-\frac{n_tkg_t\left(x_t+x_{t-1}\right)}{2} \tag{5}$$

the average welfare coming from the provision of public goods at the tier-t level. More precisely, the second term in (5) is the average per capita access cost in a tier-t jurisdiction and the last term is the per capita production cost of its administration. Indeed, since there are n_t jurisdictions at tier t, the proportion of citizens resorting to one of each of these jurisdictions is $1/n_t$. Hence, to satisfy its budget constraints, the government at tier t must levy a per capita tax equal to $\tau_t=n_tC_t$. Also, for the range of good produced by the jurisdiction we have $\int_{x_{t-1}}^{x_t}xdx=(x_t^2-x_{t-1}^2)/2$ while the average distance to the administration is $2\int_0^{1/(2n_t)} idi = 1/(4n_t^2)$, hence the expression of the average per capita access cost.

Differentiating (5) with respect to n_t, we obtain

$$n_t\sqrt{\frac{g_t}{x_t - x_{t-1}}} = \sqrt{\frac{\alpha}{4k}} \tag{6}$$

which highlights the arbitrage at each tier level between the number of jurisdictions, their scope, and the capacity of their administration. It is easily shown that this number equalizes the costs of access and provision. If $x_t - x_{t-1} = g_t = 1$, we obtain that n_t is equal to the optimal number of jurisdictions in the one-tier analysis of AS, $\sqrt{\alpha/4k}$.[13] Because the average access cost decreases and the production cost increases in (5) when n_t increases, this value corresponds to the optimal arbitrage in that case. In our setup, these costs also depend on the administration scope and capacity. Differentiating (5) with respect to g_t, equalizing to 0 and using (6) gives

$$g_t = (x_t - x_{t-1})\left(\frac{4\beta}{\sqrt{\alpha k}(x_t + x_{t-1})}\right)^{\frac{2}{1-2\beta}} \tag{7}$$

and thus

$$n_t = \left(\frac{\alpha^{1-\beta}k^{\beta}}{2^{3-2\beta}\beta}\right)^{\frac{1}{1-2\beta}} (x_t + x_{t-1})^{\frac{1}{1-2\beta}} \tag{8}$$

which correspond to optimal levels provided that $\beta \leq 1/2$.[14] Drawing a parallel with the analysis of AS by restricting the number of tiers to 1 ($t = T = 0$), with $x_0 = \bar{x} = 2$ to normalize the public task to 1, we obtain that $g_0 = \left(4\beta/(3\sqrt{\alpha k})\right)^{\frac{2}{1-2\beta}}$ which diminishes with k and α and increases with β, and $n_0 = \left(3\alpha^{1-\beta}k^{\beta}/(2^{3-2\beta}\beta)\right)^{\frac{1}{1-2\beta}}$ that increases with k and α (and from (6), we also deduce that it diminishes with β). Hence, contrary to AS, n_0 increases with k. This is due to the fact that the average per capita access cost diminishes rapidly with n_t (its derivative is proportional to n_t^{-2}) while the production cost increases linearly with both n_t and g_t. Hence, when it is possible to adjust the

[13] Of course, the number of jurisdictions (and the number of tiers T) are discrete variables, and this value corresponds to an approximation that allows us to perform comparative static exercises. More generally, the social planner's objective (4) corresponds to a mixed-integer programming problem that is solved using algorithmic methods rather than differential calculus. In order to characterize the optimal AG of this section and of the following, we consider the "relaxations" of the corresponding mixed-integer programs, i.e. problems where all variables are treated as continuous variables.

[14] This is due to the additive separability between the administration activity and money of the citizen's utility function. The first-order condition with respect to g_t equalizes the production cost to βu_t, while the one with respect to n_t equalizes the production and the access costs, leading to a net utility equal to $(1 - 2\beta)u_t$. Hence, congestion costs cannot be too large, i.e. we must have $\beta \leq 1/2$.

capacity of the administrations, an increase in the access cost or in the production cost parameter results in a downsizing of the administrations' capacity that is compensated by an increase in the number of jurisdictions. In other words, to remain at an optimal welfare level when costs are increased exogenously, the capacity of public facilities is reduced but the proximity of the government to the citizen is improved by a larger number of jurisdictions.

By limiting government to one level of administration ($T = 0$), the whole set of public goods is produced by the central government of a country, and $n_0 \geq 2$ means that the territory must be separated into several independent states. By relaxing this constraint, i.e. by allowing countries to have several levels of jurisdictions, it is possible to increase the proximity of government to citizens by dividing the public task between the jurisdictional levels of a nation, with the least accessible services being produced by local administrations. Therefore, the term $x_t + x_{t-1}$ in (8), which is equal to $x_0 + x_{-1} = \bar{x} + 1$ when $T = 0$, is possibly reduced (if $T > 0$ and thus $x_0 < \bar{x}$ at the optimum of the social planner's program), which diminishes the desirable number of countries. The optimal sharing of task between tiers satisfies $\partial W_t / \partial x_t = \partial W_{t+1} / \partial x_t$ for all $t = 0, \ldots, T-1$, i.e. the marginal welfare gain from an increase in services provided by tier t should be equal to the marginal welfare loss resulting from the induced decrease in services at level $t+1$. As apparent in (7) and (8), the administration's capacity and the number of jurisdictions at the optimum are functions of the optimal sharing of tasks between tiers $\{1, x_0, \ldots, x_t, \ldots, \bar{x}\}$. Problem (4) is thus solved by determining the optimal apportionment of tasks. It turns out that it is convenient to express x_t recursively, i.e. $x_t = \lambda_t x_{t-1}$, and to look for the optimal sequence of magnification parameters $\boldsymbol{\lambda_c} \equiv \{\lambda_{c0}, \ldots, \lambda_{cT}\}$, with $\lambda_{ct} > 1$ for all $t \in \{0, \ldots, T-1\}$ and $\lambda_{cT} = \bar{x} / \prod_{t=0}^{T-1} \lambda_{ct}$. We obtain the following results

Lemma 2 *When $\beta = 1/4$, the sequence $\boldsymbol{\lambda_c}$ is degenerate, i.e. $\lambda_{ct} = \lambda_c$ for all $t < T$, with λ_c and T solutions of*

$$\max_{\lambda_c, T} \left\{ \frac{1}{2\sqrt{\alpha k}} \left(T \frac{\lambda_c - 1}{\lambda_c + 1} + \frac{\bar{x} - \lambda_c^T}{\bar{x} + \lambda_c^T} \right) - \nu T : \lambda_c^{T+1} \geq \bar{x} > \lambda_c^T \right\} \tag{9}$$

provided that ν, α and k are not too large. In that case λ_c increases with ν, α and k and we must have $k\alpha^3 < 1/4$ for $n_0 \leq 1$, and $n_0 \geq 2$ if $\nu\sqrt{\alpha k} \geq (1 - (k\alpha^3)^{1/4})^2/2$.
When $\beta \neq 1/4$, the sequence $\boldsymbol{\lambda_c}$ is defined recursively. If β is close enough to $1/4$ so that optimal T is unaffected, $\boldsymbol{\lambda_c}$ is increasing, i.e. $\lambda_{ct} > \lambda_{ct-1}$ for $t = 1, \ldots, T-1$, with first term $\lambda_{0c} < \lambda_c$ if $\beta > 1/4$ and $\alpha k < 1$, and decreasing with first term $\lambda_{0c} > \lambda_c$ if

$\beta < 1/4$ *and* $\alpha k < 1$.

Not surprisingly, the access cost and production parameters must be small ($k\alpha^3 < 1/4$) for a single country to rule the entire territory. It is shown in the Appendix that the sequence of magnification parameters $\boldsymbol{\lambda_c}$ is defined recursively from its first term, λ_{c0}. In the case $\beta = 1/4$, the recursion is degenerate, i.e. $\lambda_{ct} = \lambda_c$ for all $t < T$, with λ_c obtained from solving (9). This reduced program reveals that the net utility is the same for each tier but the last one, given by $(2\sqrt{\alpha k})^{-1}(\lambda_c - 1)/(\lambda_c + 1)$. The optimal λ_c increases when one of the cost parameters, α, k or ν, is exogenously raised. For $\beta \neq 1/4$, magnification parameters are no longer constant. The corresponding program is given in the appendix and it is shown that when β is close to 1/4, they are approximately given by $\lambda_{ct} \approx \lambda_c + (4\beta - 1)(\Delta + t\gamma)$ where $\gamma > 0$ and $\Delta < 0$ when $\alpha k < 1$, and $\Delta > 0$ when $\alpha k > 1$. Hence, assuming that the cost parameters are not too large, this corresponds to a spreading of the sequence terms when β is larger than 1/4, with a small initial value ($\lambda_{c0} < \lambda_c$), hence a reduced scope for the central government compared to $\beta = 1/4$, and scopes increasingly wider for sub-national levels since the magnification parameters λ_{ct} are increasing for all $t < T$. Indeed, comparing scopes between to successive levels, we get

$$\frac{x_{t+1} - x_t}{x_t - x_{t-1}} = \left(1 + \frac{\lambda_{ct+1} - \lambda_{ct}}{\lambda_{ct} - 1}\right)\lambda_{ct} \approx \left(1 + \frac{(4\beta - 1)\gamma}{\lambda_c - 1}\right)\lambda_{ct}$$

neglecting second-order terms. It goes the other way around for low values of β, i.e. a central state with a large scope, and ratios of scopes that are decreasing going down the AG (as depicted Fig. 1). The next proposition spells out some other important properties of the optimal AG:

Proposition 1 *The degree of territorial dispersion, relative per capita administration's capacity and relative level of utility satisfy*

$$\frac{n_{t+1}}{n_t} = \left(1 + \frac{\lambda_{ct+1} - \lambda_{ct}}{\lambda_{ct} + 1}\right)^{1/\kappa} \lambda_{ct}^{1/\kappa},$$

$$\frac{n_t g_t}{n_{t+1} g_{t+1}} = \lambda_t^{2\beta/\kappa}\left(1 - \frac{\lambda_{t+1} - \lambda_t}{\lambda_{t+1} - 1}\right)\left(1 + \frac{\lambda_{t+1} - \lambda_t}{\lambda_t + 1}\right)^{1/\kappa}$$

and

$$\frac{u_{t+1}}{u_t} = \left(1 + \frac{\lambda_{ct+1} - \lambda_{ct}}{\lambda_{ct} - 1}\right)\left(1 - \frac{\lambda_{ct+1} - \lambda_{ct}}{\lambda_{ct+1} + 1}\right)^{2\beta/\kappa} \lambda_t^{(1-4\beta)/\kappa},$$

for $t = 1, \ldots, T - 2$, *where* $\kappa = 1 - 2\beta > 0$.

When β is close to $1/4$, proceeding down the AG to the penultimate level
i/ the number of jurisdictions increases, at an increasing (decreasing) rate when $\beta > 1/4$ ($\beta < 1/4$)
ii/ the degree of jurisdictional dispersion is stable if $\beta = 1/4$, increases if $\beta > 1/4$ and decreases if $\beta < 1/4$.
iii/ the per capita administration's capacity decreases, at an increasing (decreasing) rate when $\beta > 1/4$ ($\beta < 1/4$)
iv/ the satisfaction that citizens derive from public activity is constant if $\beta = 1/4$, increases if $\beta < 1/4$ and decreases if $\beta > 1/4$.

The above formulas simplify nicely when $\beta = 1/4$: the degree of jurisdictional dispersion in tier t, defined as n_{t+1}/n_t, i.e. the number of jurisdictions in tier $t + 1$ which geographically belong to the same jurisdiction in tier t, is given by λ_c^2 for all $t < T$, and the capacity of the administration is the largest at the central level (tier 0) and decreases at rate λ_c^3 as one proceeds down the AG. Hence, the administration at the bottom-most tier, the one the closest to its citizenry (usually, a town), has the smallest administration's capacity although it produces public goods in the upper range of the cost parameter, i.e. the more costly in terms of production and accessibility for the citizen (like garbage collection or elementary education). At the other extreme, the central government has the largest administration. Both the fraction of the population belonging to a jurisdiction (given by $1/n_t$) and the administration's capacity decrease with moving down the AG. Comparing the administration's capacity per capita at different levels, we get $g_t n_t = g_{t-1} n_{t-1}/\lambda_c$, hence capacities that also decrease down the AG. The citizens' satisfaction u_t is the same for all tiers (but the last one), as are per capita production and average access costs, which represent $2\beta = 50\%$ of the utility. Prop. 1 also shows that these general features about the number of jurisdictions and the capacity of their administrations do not change much when β is different from $1/4$; only the rate of decrease or increase are affected, depending on the case at hand. However, relative jurisdictions scopes are affected by β, as well as the relative citizen satisfaction: when β is large ($\beta > 1/4$), citizens are more satisfied with the central government than they are with sub-national levels, and the reverse otherwise. Since costs are proportional to utility, the production and average access costs at the top tier are also lower than those of sub-national levels.

As noted above for $\beta = 1/4$, a rise in the cost parameters α, k or ν, increases λ_c and thus the number of jurisdictions per tier and their range of services. The increase in n_t is reminiscent of the discussion regarding (6) which increases with α and k. The increase in scopes indicates that when these costs are very large, the territory is divided into a large number of countries with only one level of government. Conversely, when these parameters are small, the number of nations is small (possibly only one country rule the entire territory if $k\alpha^3 < 1/4$), but their governments have several levels of jurisdictions.

4 Decentralization

So far, we have considered jurisdictions that are managed by social planner appointees with no leeway to pursue their own political agendas. We now consider that at each level, (elected) governments decide on their own jurisdiction's range of services and administration's capacity. Every jurisdiction thus has the possibility of doing what it considers best for its citizens, which may not be perfectly aligned with the social planner's view. This could be because the jurisdictional authorities are fiscally responsible but do not have a say on the decisions of the other tiers' jurisdictions to which their constituents also belong, or simply because the decisions of these levels concern either a wider group of citizens (the higher jurisdictions in the AG), or subsets of their constituents (the lower jurisdictions in the AG). Politicians seeking reelection have a vested interest in focusing on the direct impact of their decisions on their own constituents, as they are identified as the sole party responsible for the choices made in their jurisdiction, and not on the decisions made by elected officials at other levels of jurisdiction.

We thus consider policymakers that must decide on the range of services and capacity of their jurisdictions to maximize the well-being of their own constituencies within the constraint of a balanced budget for the jurisdiction for which they are responsible. The resulting objective is given by (5) for a jurisdiction of tier t. As observed above, without congestion effects ($\beta = 0$), as the citizen's utility is additively separable with respect to the public goods, re-assignment of public tasks between tiers do not affect the welfare as long as the tiers' scopes do not overlap. This is not the case when the operating capacity of the administration matters since the capacity of each administration should be adjusted to its scope. As shown in the following, the entire AG is affected depending on the way the public task is apportioned between tiers. We consider a bottom-up apportionment procedure, dubbed "decentralization", that is illustrative of the recent evolution in which states are involved.

More precisely, decentralization is introduced into the AG problem as a multi-stage delegation game where the social planner first determines the number of countries and jurisdictions in each tier level (n_t for each potential t), and then delegates to their decision-makers the choice of the range of services and the capacity of their administration in an decreasing order of precedence along the AG. Delegation to decision-makers is operated successively from tier 0 to the last tier: the planner first asks decision-makers of tier 0 to choose g_0 and x_0. This policy maximizes (5) where $t = 0$, $x_{-1} = 1$. Then, the planner asks the decision-makers of tier $t = 1$ to do the same for their respective constituents under the constraint that the lower bound of their scope of services, x_0, corresponds to the choice made by tier $t = 0$ to avoid that their services overlap. Given their constituents, they all maximize (5) where $t = 1$, $x_{t-1} = x_0$ with respect to x_1 and g_1. The planner then asks the decision-makers of tier $t = 2$ to do the same for their respective citizenry under the constraint that the lower bound of their jurisdiction's range of services corresponds to the choice of the tier $t = 1$ decision-maker that rules the jurisdiction above them. The process continues until the last tier, $t = T$, which range is given by $[x_{T-1}, \bar{x}]$ and thus has only to determine its administration's capacity g_T.[15]

Before deriving the resulting AG, it is interesting to determine how a jurisdiction with objective (5) would modify the allocation of services or the capacity of its administration compared to the first-best levels if it is given the choice.

Proposition 2 *If assigned their first-best scope, jurisdictions of tier-t would choose the same administration's capacity as the social planner. However, if assigned their first-best administration's capacity, these jurisdictions have an incentive to carry out the services assigned to upper tiers in the AG, and to lower tiers too if* $\beta \leq (1/\lambda_{ct}+1)/4$.

Not surprisingly, if decision-makers are constrained in their range of services, they would not change the capacity of their administration compared to the first-best ones, since their objective in this case coincides with that of the social planner. The changes to the range of services mentioned in Prop. 2 come from the fact that the social planner compares the benefits of allocating the production of goods to jurisdictions at two consecutive levels to determine the best allocation, whereas local decision-makers only consider the effects on their own jurisdiction. Given the first-best AG, it is always beneficial for a local decision-maker to broaden the range of services of his administration

[15] We assume that jurisdictions do not coordinate either between themselves (horizontally) or with other tiers (vertically). Since the citizens' utility is additive separable between tiers, each local decision-maker adopts a dominant strategy choosing the scope and capacity of her administration's when it is her turn to make a decision.

by taking over the production of low-cost public goods. Indeed, by increasing its scope, a jurisdiction increases the citizen satisfaction by $(1-\beta)/(x_t - x_{t-1})$ percent. It also reduces the production cost of its administration by $n_t k g_t/2$, but increases the average access cost by $\alpha x_{t-1}/(4n_t)$. Hence, the total marginal benefit is

$$\frac{n_t k g_t}{2} - \frac{\alpha x_{t-1}}{4n_t} + \frac{(1-\beta)u_t}{x_t - x_{t-1}}. \tag{10}$$

At the social optimum, the production and access costs of the jurisdictions are equalized, and correspond to β percent of the gross utility provided by the jurisdiction: we have $n_t k g_t(x_t + x_{t-1})/2 = \alpha\left(x_t^2 - x_{t-1}^2\right)/(8n_t) = \beta u_t$. Using theses equalities to re-express the marginal costs, we arrive at an overall marginal effect of a decrease in x_{t-1} equal to $u_t[x_t - (4\beta-1)x_{t-1}]/(x_t^2 - x_{t-1}^2)$, which is positive since $\beta < 1/2$ and $x_t > x_{t-1}$. If a jurisdiction broadens the scope of its administration by taking over the production of goods assigned to lower jurisdictions in the AG (an increase in x_t), it suffers a raise in both the production and the access cost. Therefore, citizen's preferences for the range of services should be sufficiently large (i.e. β relatively low), for local decision-makers to find it beneficial: the marginal benefit is

$$-\frac{n_t k g_t}{2} - \frac{\alpha x_t}{4n_t} + \frac{(1-\beta)u_t}{x_t - x_{t-1}}, \tag{11}$$

which, re-expressed in utils, leads to the condition $(1-4\beta)x_t + x_{t-1} > 0$.

A similar trade-off characterizes decentralized governments, where jurisdictions adapt their administration's capacity to the scope they wish to adopt. Their capacity choice follows the same rule as that applied by the social planner, i.e. optimal capacity must equalize the marginal production cost with the marginal increase in the citizen welfare. It leads to a production cost that is equal to β percent of gross utility. However, the choice of scope is not in line with that of the social planner. Indeed, the socially optimal scope is such that the marginal benefit (11) of an increase in x_t at level t is equal to the marginal loss at level $t+1$ (the equivalent of (10) but with t replaced by $t+1$). When the local authority chooses its scope under decentralization, (11) is equal to 0. Hence, local decision-makers consider that the capacity of their administration is large enough to provide more services than those assigned by the social planner. They thus tend to oversupply given their administration's capacity. To limit the adverse effects of these choices on the overall welfare, the social planner modifies the AG by adapting the number of tiers and jurisdictions per tier.

More precisely, the delegation game under decentralization results in a AG that solves

$$\max_{T,\{n_t\}_{t=0}^{T}} \left\{ \sum_{t=0}^{T} \max_{x_t, g_t} \{W(x_{t-1}, x_t, g_t, n_t) | x_{t-1}, n_t\} - \nu T : x_{-1} = 1, x_T = \bar{x} \right\}. \tag{12}$$

As for the first-best case, it is convenient to express x_t recursively to solve problem (12) and to look for the optimal sequence of magnification parameters $\boldsymbol{\lambda_d} \equiv \{\lambda_{d0}, \ldots, \lambda_{dT}\}$, with $\lambda_{dt} > 1$ for all $t \in \{0, \ldots, T-1\}$ and $\lambda_{dT} = \bar{x} / \prod_{t=0}^{T-1} \lambda_{dt}$. Likewise, this sequence is generally defined by a recursive equation which simplifies in the case $\beta = 1/4$. More precisely,

Lemma 3 *When $\beta = 1/4$, the sequence $\boldsymbol{\lambda_d}$ is degenerate, i.e. $\lambda_{dt} = \lambda_d$ for all $t < T$, $\lambda_{dT} = \bar{x}/\lambda_d^T$, and λ_d and T are solution of*

$$\max_{\lambda_d, T} \left\{ \frac{1}{2\sqrt{\alpha k}} \left(T\Phi_d(\lambda_d) \frac{\lambda_d - 1}{\lambda_d + 1} + \frac{\bar{x} - \lambda_d^T}{\bar{x} + \lambda_d^T} \right) - \nu T : \lambda_d^{T+1} \geq \bar{x} > \lambda_d^T \right\} \tag{13}$$

where $\Phi_d(\lambda) \equiv (\lambda - 1)(\lambda + 2)^{1/2} / \lambda^{3/2}$, provided that ν, α and k are not too large. In that case $\lambda_d > 2$ and increases with ν, α and k. We must have $k\alpha^3 < 4.10^{-4}$ to have $n_0 \leq 1$, and a sufficient condition for $n_0 \geq 2$ is $k\alpha^3 \geq 2/81 \approx 15.10^{-4}$.

Due to the weight $\Phi_d(\lambda) < 1$ that affects the utility of each level $t < T$ in (13) compared to (9), the magnification parameter λ_d is large, larger than 2 in any case, and the condition to have only one state managing the entire territory is far more stringent than under the first-best AG.[16] It is shown in the appendix that the ratios of scopes, number of jurisdictions, administration capacities and citizen utilities have the same expression as in the first-best case when $\beta = 1/4$ (albeit, these expressions are more intricate in general) and that $\lambda_d > \lambda_c$. Hence, comparing the AG under the 2 regimes when $\beta \approx 1/4$, we obtain

Proposition 3 *Compared to the first-best AG, the optimal AG under decentralization entails a larger territorial dispersion, less tiers, with jurisdictions that have increased ranges of services.*

To illustrate these theoretical results, a numerical example is given Table 1 using $\alpha = .25, k = 5$ (hence $k\alpha^3 = .016$), $\nu = .01$, and $\bar{x} = 100$. This table presents the solutions to the maximization of programs (9) and (13). The resulting magnification parameters are $\lambda_c = 1.94$ and $\lambda_d = 4.07$, that correspond to jurisdictional dispersions

[16]This is thus reminiscent of AS's result that democracy leads to too many nations.

of approximately 4 and 16 respectively (we have $n_{t+1}/n_t = \lambda^2$ when $\beta = 1/4$). Columns $\hat{n}_{ct}$ and $\hat{n}_{dt}$ correspond to the rounded values of the programs' results n_{ct} and n_{dt}, and columns $\hat{W}_{ct}$ and $\hat{W}_{dt}$ to the net welfare levels computed with these rounded values. Overall, the total estimated welfare with these rounded values, $\sum_t \hat{W}_{ct} = .93$ and $\sum_t \hat{W}_{dt} = .76$, are close to the welfare levels of the corresponding program, $\sum_t W_{ct} = 1$ and $\sum_t W_{dt} = .8$.[17] Comparing the AGs, there is only one country ruling the entire territory under the first-best AG (i.e. $n_{c0} = 1$) with 6 sub-national tiers, and 2 states under decentralization with only 3 sub-national levels.

5 Eliciting citizen preferences

In this section, we illustrate how the equilibrium conditions derived above can be used to estimate the preference parameter β from a data set with standard econometric procedures. Using $x_t = x_{t-1}\lambda_{jt}$ and $g_t \approx g_{t-1}\lambda_{jt-1}^{-(1+2\beta)/(1-2\beta)}$ when β is close to 1/4, $j \in \{c, d\}$, we obtain that the ratio of production costs of two consecutive tiers satisfies

$$\frac{C_{t+1}}{C_t} = \frac{kg_{t+1}(x_{t+1} + x_t)/2}{kg_t(x_t + x_{t-1})/2} = \frac{g_{t+1}}{g_t}\lambda_t \left(1 + \frac{\lambda_{t+1} - \lambda_t}{\lambda_t + 1}\right) \approx \lambda_j^{-4\beta/(1-2\beta)}$$

when $\beta \approx 1/4$. Similarly, we have $n_t \approx n_{t-1}\lambda_j^{1/(1-2\beta)}$ when $\beta \approx 1/4$. Taking the logarithm of both expressions and eliminating the term involving $\ln \lambda_j$ on both sides, we get

$$\ln (C_t/C_{t+1}) \approx 4\beta \ln(n_{t+1}/n_t) \tag{14}$$

which holds for any AG, hence whatever the organization of the country. Using data on jurisdiction spending, it is possible to compute $Y_{z_{t+1},d} \equiv \ln(C_{z_t,d}/C_{z_{t+1},d})$ for each jurisdiction $z_{t+1} \in \{1, \ldots, n_{t+1}\}$ belonging to tier-$t+1$ located below jurisdiction z_t in the AG at each date d. As $1/n_t$ corresponds to the share of the population of a tier-t jurisdiction, $n_{z_{t+1},d}/n_{z_t,d}$ corresponds to the inverse of the share of the z_t citizenry that belongs to jurisdiction z_{t+1} at date d. These individual characteristics, together with other observable characteristics (e.g., disposable income, population density), can be used to estimate the coefficient β using standard panel data procedures.

We apply this methodology on annual federal and state expenditures in the United States over the period 1977-2015 collected by the Tax Policy Center[18] to estimate the

[17] Note that the estimated welfare levels $\sum_t \hat{W}_{jt}$ are obtained by rounding off the numbers of jurisdictions per tier, which do not lead to perfectly nested tiers (n_{t+1}/n_t should also be integer values). Adjusting for this constraint would reduced the welfare levels.

[18] Details on the Tax Policy Center are available on its website `https://www.urban.org/`

following relationships:

$$Y_{z_1,d} = \beta.P_{z_1,d} + a.X_{z_1,d} + \varphi_{z_1} + \delta_d + \varepsilon_{z_1,d} \tag{15}$$

$$Y_{z_1,d} = \sum_{\tau=1}^{4} \beta_\tau \mathbb{1}_{\{z_1\in\tau\}} P_{z_1,d} + a.X_{z_1,d} + \varphi_{z_1} + \delta_d + \varepsilon_{z_1,d} \tag{16}$$

$$Y_{z_1,d} = \sum_{Z\in\{A,B\}} \beta_Z \mathbb{1}_{\{z_1\in Z\}} P_{z_1,d} + a.X_{z_1,d} + \varphi_{z_1} + \delta_d + \varepsilon_{z_1,d} \tag{17}$$

where z_1 indexes the states ($z_1 \in \{1, \ldots, 50\}$) and $d = 1977, \ldots, 2015$. Here, $Y_{z_1,d}$ corresponds to the log of the date-d ratio of the federal expenditures (net of grants to state and local jurisdictions) over state z_1 government spending, $P_{z_1,d} = 4\ln(pop_{0,d}/pop_{z_1,d})$ accounts for the corresponding population share, $X_{z_1,d}$ is a vector of time-varying control variables composed of the income per capita, state population density and the partisan composition of the state legislatures. The income per capita captures the demand for public goods and services, which is expected to increase state expenditures (see for instance Borcherding & Deacon, 1972, and Ladd, 1992, for estimations on U.S. data), and therefore to reduce $Y_{z_1,d}$ in our setting. We also consider the partisan composition of state legislatures, as the control of both chambers by democrats is usually shown to lead to significantly higher state expenditures per capita.[19] The variable φ_{z_1} corresponds to state-specific effects that capture time-invariant unobserved heterogeneity, δ_d are year fixed effects accounting for the influence of variables affecting all states identically in year d, and $\varepsilon_{z_1,d}$ is the error term. Table 2 presents the summary statistics of all variables used in our regressions.

Equations (15)–(17) differ by the way β is estimated. In (15), we suppose that citizens preferences are the same whatever the state, while (16) allows for a heterogeneity among the four regional divisions used by the Census Bureau (Midwest, Northeast, South, West), that we denote by $\tau = 1, ..., 4$. Accordingly, $\mathbb{1}_{\{z_1\in\tau\}}$ is a dummy variable equal to 1 if the state z_1 belongs to region τ, and 0 otherwise. Therefore, $\hat{\beta}_1$ is the measure of β for the Midwest, $\hat{\beta}_2$ the one corresponding to the Northeast, $\hat{\beta}_3$ for the South and $\hat{\beta}_4$ for the West. The last specification, (17), allows us to test for a coarser partition of the states into two groups, denoted by A (Northeast & West) and B (Midwest & South) with $\mathbb{1}_{\{z_1\in A\}}$ and $\mathbb{1}_{\{z_1\in B\}}$ being the corresponding dummy variables. Therefore, $\hat{\beta}_A$ is the measure of β for the West and Northeast regions taken together,

---

`policy-centers/urban-brookings-tax-policy-center`. Queries on its dataset can be performed at `http://slfdqs.taxpolicycenter.org/index.cfm`.

[19]See Besley & Case (2003) for a survey of the numerous empirical investigations of the role of political institutions in the United States.

and $\hat{\beta}_B$ for the South and Midwest regions.

For the three specifications (15)–(17), we control for the correlation of the error term over time at the state level using cluster-robust standard errors in our panel fixed effects estimations. Table 3 gives the estimation results.

The value of β estimated for the whole country is 0.18 (columns (1) and (2); the differences between them being because the state legislature variables are omitted in the first one)[20], which belongs to the interval (0,1/2) required by our theoretical model to ensure a positive welfare. Observe that none of the estimates corresponding to the partisan composition of state legislatures and the other characteristics (income and density) are statistically significant in the regressions. These results are corroborated by a Pew Research Center's survey, reported in Table 4 (Pew Research Center, 2013). It shows that whatever their partisanship, individuals express a more favorable view of their local government than their state government, and that the federal government in Washington earns the lowest percentage points of favorable opinions.

Disaggregating β over the four U.S. regions (column (3) of Table 3) reveals a partition of the U.S. territory. Indeed, we obtain that this coefficient is lower in the Northeast and West (0.182 and 0.147) than in the Midwest and South (0.279 and 0.238). Tests for equality of the estimated regions' βs, presented in Table 5, confirm that the pairwise equalities $\hat{\beta}_1 = \hat{\beta}_2$ (Midwest and Northeast), $\hat{\beta}_1 = \hat{\beta}_4$ (Midwest and West) and $\hat{\beta}_3 = \hat{\beta}_4$ (South and West) are rejected at the 10% significance level (even at the 5% level for the latter equality).

Column (4) of Table 3 gives the estimates of this coefficient when the regions are grouped into two zones, Northeast & West (zone A), and Midwest & South (zone B). They confirm a value much larger for Midwest & South than for Northeast & West (0.251 versus 0.153). The last column of Table 5 shows that the equality of these estimates is rejected at the 5% confidence level. Hence, from Prop. 1, we may expect that on average, citizens in the Northeast and West are more appreciative of state-provided services than those produced by the federal government, while citizens in the Midwest and South are more likely to be equally satisfied with both levels of government.

[20]As a consequence, Nebraska is included only in regression (1).

6 Conclusion

The problem of organizing a country's government entails both horizontal and vertical dimensions and raises the question of the allocation of public services over the tiers. Such an organization depends on citizens' expectation about the quality of public services. We propose a simple model that allows us to characterize and compare the desirable features of the first-best and the decentralized AGs. Our approach offers a first theoretical foundation for multi-level administrative structures based on the trade-off between concentrating the production of public goods, so as to achieve economies of scope, and limiting congestion effects, in order to improve the quality of public services as experienced by citizens. It highlights the differences between the social optimum and the result of having autonomous jurisdictions. Indeed, decision-makers, whether at the central tier or at sub-national tiers, have an incentive to increase the range of their services with respect to the first-best levels. The results we obtain also allow us to perform structural empirical investigations, as illustrated by the elicitation methods of the citizen preference parameter that we detail and apply using U.S. data. This work could be extended on both fronts. From an empirical perspective, organizational choices made by a country that has decided to give more leeway to decision-makers should reveal citizens preferences: if they value far more the overall range of services produced by the administrations than their ability to better perform their tasks, scopes of tiers should increase and the number of bureaucratic layers should decrease. From a theoretical perspective, we assume that jurisdictions have their budget balanced, i.e. that local decision makers are fiscally responsible, and we rule out transfers among jurisdictional tiers of the same country. By relaxing these assumptions, one could investigate how strategic interactions, from lower jurisdictions to attract more grants or bailouts of higher tiers, or from higher jurisdictions to give incentives to lower tiers, affect the decentralized architecture of government. Also, we suppose that the central planner has perfect information about the citizens preferences. Information asymmetry, with sub-national decision-makers being better informed about their constituency than their higher level counterparts, could impact the organization of the government. Another avenue of research is to allow for a heterogeneous distribution of the population in order to relax the perfect symmetry of our framework and thus modeling the growing trend of asymmetric decentralization, i.e., when same-tiers jurisdictions have different competencies or tax powers.

References

Alesina, A. & Spolaore, E. (1997). On the number and size of nations. *The Quarterly Journal of Economics*, 112(4), 1027–1056.

Bardhan, P. K. & Mookherjee, D. (2006). *Decentralization and local governance in developing countries: A comparative perspective*. Mit Press.

Besley, T. & Case, A. (2003). Political institutions and policy choices: Evidence from the united states. *Journal of Economic Literature*, 41(1), 7–73.

Bogomolnaia, A., Le Breton, M., Savvateev, A., & Weber, S. (2006). Stability under unanimous consent, free mobility and core. *Int J Game Theory*, 35(2), 185–204.

Bogomolnaia, A., Le Breton, M., Savvateev, A., & Weber, S. (2008). Heterogeneity gap in stable jurisdiction structures. *Journal of Public Economic Theory*, 10(3), 455–473.

Bolton, P. & Roland, G. (1997). The breakup of nations: A political economy analysis. *The Quarterly Journal of Economics*, 112(4), 1057–1090.

Borcherding, T. E. & Deacon, R. (1972). The demand for the services of non-federal governments. *American Economic Review*, 62(5), 891–901.

Caliendo, L. & Rossi-Hansberg, E. (2012). The impact of trade on organization and productivity. *The quarterly journal of economics*, 127(3), 1393–1467.

Calvo, G. A. & Wellisz, S. (1978). Supervision, loss of control, and the optimum size of the firm. *Journal of political Economy*, 86(5), 943–952.

Casella, A. (2001). The role of market size in the formation of jurisdictions. *The Review of Economic Studies*, 68(1), 83–108.

Chen, C. (2017). Management quality and firm hierarchy in industry equilibrium. *American Economic Journal: Microeconomics*, 9(4), 203–44.

Demange, G. (1994). Intermediate preferences and stable coalition structures. *Journal of Mathematical Economics*, 23, 45–58.

Dexia (2007). Subnational public finance in the european union. *Dexia Economic Outlook*, Trends 2000 / 2006.

Garicano, L. (2000). Hierarchies and the organization of knowledge in production. *Journal of political economy*, 108(5), 874–904.

Geanakoplos, J. & Milgrom, P. (1991). A theory of hierarchies based on limited managerial attention. *Journal of the Japanese and international Economies*, 5(3), 205–225.

Greenberg, J. & Weber, S. (1986). Strong tiebout equilibrium under restricted preferences domain. *Journal of Economic Theory*, 38, 101–117.

Guesnerie, R. & Oddou, C. (1981). Second best taxation as a game. *Journal of Economic Theory*, 25(1), 67–91.

Hochman, O., Pines, D., & Thisse, J. F. (1995). On the optimal structure of local governments. *The American Economic Review*, 85(5), 1224–1240.

Hotelling, H. (1929). Stability in competition. *The Economic Journal*, 39(153), 41–57.

Janeba, E. & Wilson, J. D. (2011). Optimal fiscal federalism in the presence of tax competition. *Journal of Public Economics*, 95(11-12), 1302–1311.

Jehiel, P. & Scotchmer, S. (2001). Constitutional rules of exclusion in jurisdiction formation. *The Review of Economic Studies*, 68(2), 393–413.

Ladd, H. (1992). Population growth, density and the costs of providing public services. *Urban Studies*, 29(2), 273–295.

OECD (2019). *Making Decentralisation Work: A Handbook for Policy-Makers, OECD Multi-level Governance Studies*. OECD Publishing, Paris.

OECD/UCLG (2019). *Report of the World Observatory on Subnational Government Finance and Investment Country Profiles*. OECD Publishing, Paris.

Panizza, U. (1999). On the determinants of fiscal centralization: Theory and evidence. *Journal of public economics*, 74(1), 97–139.

Pew Research Center (April 15, 2013). State governments viewed favorably as federal rating hits new low. Pew Research Center, Washington, D.C. `http://www.people-press.org/category/publications/2013/page/4/` (as at August 2018).

Qian, Y. (1994). Incentives and loss of control in an optimal hierarchy. *The Review of Economic Studies*, 61(3), 527–544.

Radner, R. (1992). Hierarchy: The economics of managing. *Journal of economic literature*, 30(3), 1382–1415.

Van Zandt, T. (1999). Real-time decentralized information processing as a model of organizations with boundedly rational agents. *The Review of Economic Studies*, 66(3), 633–658.

Van Zandt, T. (2003). Real-time hierarchical resource allocation with quadratic costs. *Available at SSRN 454140*.

Wilson, J. D. & Janeba, E. (2005). Decentralization and international tax competition. *Journal of Public Economics*, 89(7), 1211–1229.

Appendix

A Proof of lemma 1

The program (3) boils down to

$$\max_{\substack{T,\\ \{x_t,n_t,\{S_{z_t},\bar{\ell}_{z_t},g_{z_t}\}\}_{t=0}^{T}}} \sum_{t=0}^{T}\sum_{z_t=1}^{n_t}\left[\left(g_{z_t}^{\beta}\left(x_t-x_{t-1}\right)^{1-\beta}-\frac{\alpha\left(x_t^2-x_{t-1}^2\right)}{2}\bar{\ell}_{z_t}\right)S_{z_t}-kg_{z_t}\frac{x_t+x_{t-1}}{2}\right]-\nu T$$

under the conditions $x_{-1}=1$, $x_T=\overline{x}$ and $\sum_{z_t=1}^{n_t}S_{z_t}=1$ for all t, where $\bar{\ell}_{z_t}$ is the average distance between the residents and the administration of jurisdiction z_t. Because individuals are located uniformly over the territory, $\bar{\ell}_{z_t}$ is minimized when the administration is located in the middle of the jurisdiction whatever the tier and the number of jurisdictions in a tier. As a consequence, the average distance to their administration is $\bar{\ell}_{z_t}=\frac{1}{S_{z_t}/2}\int_0^{S_{z_t}/2}\ell d\ell=S_{z_t}/4$. We can thus rewrite the program as

$$\max_{\substack{T,\\ \{x_t,n_t,\{S_{z_t},\bar{\ell}_{z_t},g_{z_t}\}\}_{t=0}^{T}}} \sum_{t=0}^{T}\sum_{z_t=1}^{n_t}\left[S_{z_t}g_{z_t}^{\beta}\left(x_t-x_{t-1}\right)^{1-\beta}-\frac{\alpha\left(x_t^2-x_{t-1}^2\right)}{8}S_{z_t}^2-kg_{z_t}\frac{x_t+x_{t-1}}{2}\right]-\nu T \tag{18}$$

under the conditions $x_{-1}=1$, $x_T=\overline{x}$ and $\sum_{z_t=1}^{n_t}S_{z_t}=1$ for all t, where $\sum_{z_t=1}^{n_t}S_{z_t}^2$ is minimized when tier t jurisdictions are of equal length: $S_{z_t}=1/n_t$ for all t. As jurisdictions are of equal length and their range of services are identical, we also have the same administration's capacity at each tier level: $g_{z_t}=g_t$ for all t. Replacing in (18) and developing gives (4).

B Proof of Lemma 2

The first-order conditions (FOCs) w.r.t. n_t and g_t lead to

$$\frac{\alpha\left(x_t^2-x_{t-1}^2\right)}{4n_t^2}-kg_t\left(x_t+x_{t-1}\right)=0, t\in\{0,\ldots,T\} \tag{19}$$

and

$$\beta u_t-n_tkg_t\left(x_t+x_{t-1}\right)/2=0, t\in\{0,\ldots,T\}, \tag{20}$$

respectively. Replacing the expression of g_t coming from (20) in the utility gives

$$u_t = \left(\frac{2\beta}{kn_t \left(x_t + x_{t-1}\right)} \right)^{\beta/(1-\beta)} \left(x_t - x_{t-1}\right), \tag{21}$$

which plugged back in (20) gives

$$g_t = \left(\frac{2\beta}{kn_t \left(x_t + x_{t-1}\right)} \right)^{1/(1-\beta)} \left(x_t - x_{t-1}\right), t \in \{0, \ldots, T\}. \tag{22}$$

From (19) and (20) we get

$$\frac{\alpha \left(x_t^2 - x_{t-1}^2\right)}{8n_t} = \frac{n_t k g_t \left(x_t + x_{t-1}\right)}{2} = \beta u_t, \tag{23}$$

and thus $W_t = u_t(1-2\beta)$ which is positive iff $\beta < 1/2$. From (19), we get

$$g_t = \frac{\alpha(x_t - x_{t-1})}{4kn_t^2}. \tag{24}$$

Identifying with (22), it comes

$$n_t = \left(x_t + x_{t-1}\right)^{1/\kappa} K \tag{25}$$

where $\kappa = 1 - 2\beta$, and $K = \left(k^\beta \alpha^{1-\beta} 2^{-(3-2\beta)}/\beta\right)^{1/\kappa}$. The FOC w.r.t. x_t for $t \in \{0, \ldots, T-1\}$ is given by

$$(1-\beta)\left(\frac{u_t}{x_t - x_{t-1}} - \frac{u_{t+1}}{x_{t+1} - x_t}\right) - \frac{\alpha x_t}{4}\left(\frac{1}{n_t} - \frac{1}{n_{t+1}}\right) - \frac{k}{2}(n_t g_t + n_{t+1} g_{t+1}) = 0. \tag{26}$$

Using (20) to substitute for $n_t g_t$ and $n_{t+1} g_{t+1}$ in the last term gives

$$(1-\beta)\left(\frac{u_t}{x_t - x_{t-1}} - \frac{u_{t+1}}{x_{t+1} - x_t}\right) - \frac{\alpha x_t}{4}\left(\frac{1}{n_t} - \frac{1}{n_{t+1}}\right) - \beta\left(\frac{u_t}{x_t + x_{t-1}} + \frac{u_{t+1}}{x_{t+1} + x_t}\right) = 0$$

for $t \in \{0, \ldots, T-1\}$. Multiplying by $n_{t+1} n_t$ to get

$$(1-\beta)\left(\frac{n_{t+1} n_t u_t}{x_t - x_{t-1}} - \frac{n_{t+1} n_t u_{t+1}}{x_{t+1} - x_t}\right) - \frac{\alpha x_t}{4}\left(n_{t+1} - n_t\right) - \beta\left(\frac{n_{t+1} n_t u_t}{x_t + x_{t-1}} + \frac{n_{t+1} n_t u_{t+1}}{x_{t+1} + x_t}\right) = 0,$$

and using (23) which gives $n_t u_t = \alpha\left(x_t^2 - x_{t-1}^2\right)/8\beta$ to substitute for $n_t u_t$ and $n_{t+1} u_{t+1}$

yields

$$(1-\beta)\left[n_{t+1}(x_t + x_{t-1}) - n_t(x_{t+1} + x_t)\right] - 2\beta x_t\left(n_{t+1} - n_t\right) - \beta\left[n_{t+1}(x_t - x_{t-1}) + n_t(x_{t+1} - x_t)\right] = 0$$

for all $t \in \{0, \dots, T-1\}$. Collecting terms, we get

$$\frac{x_t(1-4\beta) + x_{t+1}}{x_t(1-4\beta) + x_{t-1}} = \frac{n_{t+1}}{n_t} = \left(\frac{x_{t+1} + x_t}{x_t + x_{t-1}}\right)^{1/\kappa}$$

using (25). Using the boundary condition $x_{-1} = 1$, x_t can be expressed as $x_t = \lambda_t x_{t-1} = \prod_{\tau=0}^{t} \lambda_\tau$ where $\lambda_t > 1$ for all $t \in \{0, \dots, T\}$ w.l.g., and the outer equality can be expressed as

$$\left(\lambda_t \frac{\lambda_{t+1} + 1}{\lambda_t + 1}\right)^{1/\kappa} = \lambda_t \frac{1 - 4\beta + \lambda_{t+1}}{\lambda_t(1-4\beta) + 1}$$

which gives

$$\frac{1 - 4\beta + \lambda_{t+1}}{(\lambda_{t+1} + 1)^{1/\kappa}} = \frac{\lambda_t^{2\beta/\kappa}[\lambda_t(1-4\beta) + 1]}{(\lambda_t + 1)^{1/\kappa}}, t \in \{0, \dots, T-1\}. \tag{27}$$

We thus have $f_1(\lambda_{t+1}) = f_2(\lambda_t)$ where $f_1(\lambda) \equiv (1 - 4\beta + \lambda)(\lambda + 1)^{-1/\kappa}$ and $f_2(\lambda) \equiv \lambda^{2\beta/\kappa}[\lambda(1-4\beta) + 1](\lambda + 1)^{-1/\kappa}$. As

$$f_1'(\lambda) = \frac{-2\beta(\lambda - 1)}{\kappa(1+\lambda)^{2(1-\beta)/\kappa}} \tag{28}$$

is negative for all $\lambda > 1$, (27) can be expressed as $\lambda_{t+1} = F(\lambda_t) \equiv f_1^{-1}(f_2(\lambda_t))$ which defines a first-order recurrence equation specifying the sub-sequence $\{\lambda_t\}_{t=1\dots,T-1}$ given λ_0: $\lambda_t = F^t(\lambda_0) \equiv F(F^{t-1}(\lambda_0))$ (with $F^0 = Id$). The last term of the sequence is deduced from the boundary condition $x_T = \bar{x}$ which gives $\lambda_T = \bar{x}/\prod_{t=0}^{T-1} F^t(\lambda_0)$. We have $\lambda_{t+1} > 1$ if $F(\lambda_t) > 1$ for $\lambda_t > 1$. As $f_1(1) = f_2(1)$, we have $F(1) = 1$, and thus $\lambda_{t+1} > 1$ if $F'(\lambda_t) \geq 0$ for all $\lambda_t > 1$. Using

$$\begin{aligned} f_2'(\lambda) &= \frac{-2\beta(\lambda - 1)}{\kappa\lambda^{(1-4\beta)/\kappa}(1+\lambda)^{2(1-\beta)/\kappa}}, \\ f_2''(\lambda) &= -\frac{2\beta}{\kappa^2}\frac{1 - 4\beta + \lambda[3 - 2\lambda + 4(\lambda - 1)\beta]}{\lambda^{2-2\beta/\kappa}(1+\lambda)^{2+1/\kappa}} \end{aligned} \tag{29}$$

and

$$f_1''(\lambda) = \frac{2\beta}{\kappa^2}\frac{-3 + \lambda + 4\beta}{(1+\lambda)^{2+1/\kappa}},$$

we obtain

$$\frac{d\lambda_{t+1}}{d\lambda_t} = \frac{f_2'(\lambda_t)}{f_1'(\lambda_{t+1})} = \lambda_t^{(4\beta-1)/\kappa} \frac{\lambda_t - 1}{\lambda_{t+1} - 1} \left(\frac{\lambda_{t+1}+1}{\lambda_t + 1}\right)^{2(1-\beta))/\kappa}$$

for $\lambda_t > 1$, and using L'hospital rule, $\lim_{\lambda_t \searrow 1} d\lambda_{t+1}/d\lambda_t = f_2''(1)/f_1''(1) = 1$. We thus have $F'(\lambda_t) > 0$ for all $\lambda_t \geq 1$, and thus $\lambda_t > 1$ for all $t \in \{1, \ldots, T-1\}$ if $\lambda_0 > 1$. In the case $\beta = 1/4$, as $f_1(\lambda) = f_2(\lambda) = \hat{f}(\lambda) \equiv \lambda/(1+\lambda)^2$, we have $\lambda_{t+1} = \lambda_t = \lambda_c$ for all $t \in \{0, \ldots, T-1\}$.

T and λ_0 are derived as follows. Using $x_t = \prod_{\tau=0}^{t} \lambda_\tau$, (25) can be expressed as

$$n_t = K\left[(\lambda_t + 1)\, x_{t-1}\right]^{1/\kappa} = K\left((\lambda_t + 1)\prod_{\tau=0}^{t-1} \lambda_\tau\right)^{1/\kappa},$$

and substituting in (21), it comes

$$u_t = \left(\frac{16\beta^2}{k\alpha\,(x_t + x_{t-1})^2}\right)^{\beta/\kappa} (x_t - x_{t-1}) = \left(\frac{16\beta^2}{k\alpha}\right)^{\beta/\kappa} \frac{(\lambda_t - 1)\prod_{\tau=0}^{t-1} \lambda_\tau^{(1-4\beta)/\kappa}}{(\lambda_t + 1)^{2\beta/\kappa}} \tag{30}$$

for $t \in \{0, \ldots, T-1\}$, and

$$u_T = \left(\frac{16\beta^2}{k\alpha\,(x_T + x_{T-1})^2}\right)^{\beta/\kappa} (x_T - x_{T-1}) = \left(\frac{16\beta^2}{k\alpha}\right)^{\beta/\kappa} \frac{\bar{x} - \prod_{\tau=0}^{T-1} \lambda_\tau}{\left(\bar{x} + \prod_{\tau=0}^{T-1} \lambda_\tau\right)^{2\beta/\kappa}} \tag{31}$$

for the last tier. The boundary condition $x_T = \bar{x}$ implies that

$$\prod_{\tau=0}^{T} F^\tau(\lambda_0) \geq \bar{x} > \prod_{\tau=0}^{T-1} F^\tau(\lambda_0), \tag{32}$$

and λ_0 and T are derived from solving $\max_{\lambda_0, T} \sum_{t=0}^{T} W_t - \nu T = \max_{\lambda_0, T} \kappa \sum_{t=0}^{T} u_t - \nu T$ using (27) and (30)–(32). In the case $\beta = 1/4$, we have $x_t = \lambda_c^{t+1}$, $W_t = u_t/2$ where

$$u_t = \frac{1}{\sqrt{\alpha k}} \frac{\lambda_c - 1}{\lambda_c + 1} \equiv u_c(\lambda_c), t \in \{0, \ldots, T-1\},$$

with $\lambda_c^{T+1} \geq \bar{x} > \lambda_c^T$. We thus have to solve

$$\max_{\lambda_c, T} \left\{ T\frac{\lambda_c - 1}{\lambda_c + 1} + \frac{\bar{x} - \lambda_c^T}{\bar{x} + \lambda_c^T} - \nu T : \lambda_c^{T+1} \geq \bar{x} > \lambda_c^T \right\} \tag{33}$$

where $\nu \equiv 2\nu\sqrt{\alpha k}$. Neglecting the constraints (and treating the discrete variable T as a continuous variable), the FOCs with respect to T and λ_c gives

$$\frac{\lambda_c - 1}{\lambda_c + 1} - \nu - \frac{2\bar{x}\lambda_c^T \ln \lambda_c}{(\bar{x} + \lambda_c^T)^2} = 0$$

and

$$\frac{2}{(\lambda_c + 1)^2} - \frac{2\bar{x}\lambda_c^{T-1}}{(\bar{x} + \lambda_c^T)^2} = 0,$$

respectively. Multiplying the last equation by $\lambda_c \ln \lambda_c$ and identifying with the first one yields

$$\frac{2\lambda_c \ln \lambda_c}{(\lambda_c + 1)^2} = \frac{2\bar{x}\lambda_c^T \ln \lambda_c}{(\bar{x} + \lambda_c^T)^2} = \frac{\lambda_c - 1}{\lambda_c + 1} - \nu,$$

hence $g_c(\lambda_c) = \nu$ where $g_c(\lambda) \equiv (\lambda^2 - 1 - 2\lambda \ln \lambda)/(\lambda + 1)^2$. As $g_c'(\lambda) = 2(\lambda - 1)\ln \lambda/(1 + \lambda)^3 > 0$ for $\lambda > 1$, $g_c(\lambda)$ is strictly increasing over $(1, +\infty)$. Hence, there is only one solution to $g_c(\lambda_c) = \nu$. Using $\ln \lambda = -\ln(1 - (\lambda - 1)/\lambda) \geq (\lambda - 1)/\lambda$, we have

$$g_c(\lambda) \leq [\lambda^2 - 1 - 2(\lambda - 1)]/(\lambda + 1)^2 = [(\lambda - 1)/(\lambda + 1)]^2 < 1,$$

implying that we must have $\nu < 1$, i.e. $\nu\sqrt{\alpha k} < 1/2$ for (33) to have a solution, and more generally we must have $\nu < g_c(\bar{x})$. Assuming this is the case, as $\sqrt{\nu} < 1$, we get using $g_c(\lambda_c) = \nu \leq [(\lambda_c - 1)/(\lambda_c + 1)]^2$ that $\lambda_c \geq (1 + \sqrt{\nu})/(1 - \sqrt{\nu})$. As $n_0 = \sqrt{k\alpha^3}\,(\lambda_c + 1)^2/2$, a sufficient condition for $n_0 \geq 2$ is $\sqrt{k\alpha^3}\,(2/(1 - \sqrt{\nu}))^2/2 \geq 2$, hence $\nu \geq (1 - (k\alpha^3)^{1/4})^2/(2\sqrt{\alpha k})$. Also, as $\lambda_c > 1$, to allow for $n_0 \leq 1$, it should be the case that $2\sqrt{k\alpha^3} < 1$, hence $k\alpha^3 < 1/4$.

For $\beta \neq 1/4$, the equality $f_1(\lambda_{t+1}) = f_2(\lambda_t)$ can be approximated when β is close to $1/4$, i.e. $\beta = 1/4 \pm \varepsilon$, $\varepsilon > 0$ small, using the first-order approximation

$$f_i(\lambda) \approx \hat{f}(\lambda_c) + (\lambda - \lambda_c) f_i'(\lambda_c)|_{\beta=1/4} + (\beta - 1/4)(df_i(\lambda_c)/d\beta)|_{\beta=1/4},$$

for $i = 1, 2$. Using (28), (29),

$$\frac{df_1(\lambda)}{d\beta} = f_1(\lambda)\left(\frac{-4}{1 - 4\beta + \lambda} - \frac{2\ln(\lambda + 1)}{\kappa^2}\right)$$

and

$$\frac{df_2(\lambda)}{d\beta} = f_2(\lambda)\left(\frac{2\ln \lambda}{\kappa^2} - \frac{4\lambda}{(1 - 4\beta)\lambda + 1} - \frac{2\ln(\lambda + 1)}{\kappa^2}\right),$$

yields

$$f_1(\lambda) \approx \hat{f}(\lambda_c)\left[1+\frac{(\lambda-\lambda_c)(1-\lambda_c)}{(1+\lambda_c)\lambda_c}-(4\beta-1)\left(\frac{1}{\lambda_c}+2\ln(\lambda_c+1)\right)\right]$$

and

$$f_2(\lambda) \approx \hat{f}(\lambda_c)\left[1+\frac{(\lambda-\lambda_c)(1-\lambda_c)}{(1+\lambda_c)\lambda_c}-(4\beta-1)(-2\ln\lambda_c+\lambda_c+2\ln(\lambda_c+1))\right].$$

Hence, (27) can be approximated by

$$0 = f_1(\lambda_{t+1}) - f_2(\lambda_t) \approx \frac{\hat{f}(\lambda_c)}{\lambda_c}\left(\frac{(\lambda_{t+1}-\lambda_t)(1-\lambda_c)}{1+\lambda_c}+(4\beta-1)h(\lambda_c)\right)$$

where $h(\lambda) = \lambda^2 - 1 - 2\lambda\ln\lambda$, implying

$$\lambda_{t+1}-\lambda_t \approx \frac{\lambda_c+1}{\lambda_c-1}(4\beta-1)h(\lambda_c). \tag{34}$$

As $h'(\lambda) = 2(\lambda - \ln\lambda - 1)$ and $h''(\lambda) = 2(1-1/\lambda) > 0$ when $\lambda > 1$, we have $h(1) = 0, h'(1) = 0$ and $h'(\lambda) > 0$ for $\lambda > 1$, implying that $h(\lambda) > 0$ for $\lambda > 1$. Consequently, $\lambda_{t+1} > \lambda_t$ iff $\beta > 1/4$, i.e. (27) defines an increasing sequence if $\beta > 1/4$, and a decreasing one if $\beta < 1/4$. Using $\ln\lambda = -\ln(1-(\lambda-1)/\lambda) \geq (\lambda-1)/\lambda$, we get $h(\lambda) \leq \lambda^2 - 1 - 2(\lambda-1) = (\lambda-1)^2$ and thus $\lambda_{t+1} - \lambda_t \leq (\lambda_c^2-1)(4\beta-1)$. The solution of (34) is given by $\lambda_t = \lambda_c + \delta(\beta) + (4\beta-1)t\gamma$, $t \in \{0, \ldots, T-1\}$, where $\gamma = h(\lambda_c)(\lambda_c+1)/(\lambda_c-1)$ and $\delta(\beta)$ a function verifying $\delta(1/4) = 0$. When β is close to 1/4, we also have $\lambda_t \approx \lambda_c + (4\beta-1)(\Delta + t\gamma) \equiv \hat{\lambda}_t$, where $\Delta = \delta'(1/4)/4$ can be derived as follow. Substituting the approximation $\hat{\lambda}_t$ of λ_t in (30) using

$$\begin{aligned}
\prod_{\tau=0}^{t-1} \hat{\lambda}_\tau^{(1-4\beta)/\kappa} &= \exp\left(\frac{1-4\beta}{1-2\beta}\sum_{\tau=0}^{t-1}\ln(\lambda_c+(4\beta-1)(\delta(\beta)+t\gamma))\right) \\
&\approx 1+\left(\beta-\frac{1}{4}\right)\frac{d}{d\beta}\left[\frac{1-4\beta}{1-2\beta}\right]_{\beta=1/4}\sum_{\tau=0}^{t-1}\ln\lambda_c \\
&= 1-2\,(4\beta-1)\,t\ln\lambda_c,
\end{aligned}$$

and

$$\begin{aligned}\left(\hat{\lambda}_t+1\right)^{2\beta/\kappa} &= \exp\left(\frac{2\beta}{1-2\beta}\ln(\lambda_c+(4\beta-1)(\Delta+t\gamma)+1)\right)\\ &\approx (\lambda_c+1)\left[1+\left(\beta-\frac{1}{4}\right)\left(\frac{d}{d\beta}\left[\frac{2\beta}{1-2\beta}\right]_{\beta=1/4}\ln(\lambda_c+1)+\frac{4(\Delta+t\gamma)}{\lambda_c+1}\right)\right]\\ &= \hat{\lambda}_t+1+2\,(4\beta-1)\,(\lambda_c+1)\ln(\lambda_c+1),\end{aligned}$$

we get

$$u_t \approx \left(\frac{16\beta^2}{k\alpha}\right)^{\beta/\kappa}\frac{\lambda_c-1-(4\beta-1)\,[2(\lambda_c-1)t\ln\lambda_c-(\Delta+t\gamma)]}{\lambda_c+1+(4\beta-1)\,[\Delta+t\gamma+2(\lambda_c+1)\ln(\lambda_c+1)]}$$

neglecting second-order terms. A linear approximation of the last term gives

$$u_t \approx \left(\frac{16\beta^2}{k\alpha}\right)^{\beta/\kappa}\left(\frac{\lambda_c-1}{\lambda_c+1}\,[1-2\,(4\beta-1)\,(t\ln\lambda_c+\ln(\lambda_c+1))]+2\,(4\beta-1)\,\frac{\Delta+t\gamma}{(\lambda_c+1)^2}\right)$$

for $t\in\{0,\ldots,T-1\}$. Summing over t, it comes

$$\begin{aligned}\sum_{t=0}^{T-1}u_t &\approx \left(\frac{16\beta^2}{k\alpha}\right)^{\beta/\kappa}\sum_{t=0}^{T-1}\left(\frac{\lambda_c-1}{\lambda_c+1}\,[1-2\,(4\beta-1)\,(t\ln\lambda_c+\ln(\lambda_c+1))]+2\,(4\beta-1)\,\frac{\Delta+t\gamma}{(\lambda_c+1)^2}\right)\\ &= \left(\frac{16\beta^2}{k\alpha}\right)^{\beta/\kappa}T\left(\frac{\lambda_c-1}{\lambda_c+1}(1-(4\beta-1)\,[(T-1)\ln\lambda_c+2\ln(\lambda_c+1)])+2\,(4\beta-1)\,\frac{\Delta+(T-1)\gamma}{(\lambda_c+1)^2}\right).\end{aligned}$$

Also, using

$$\begin{aligned}x_t \approx \prod_{\tau=0}^{t}\hat{\lambda}_\tau &= \exp\left(\sum_{\tau=0}^{t}\ln(\lambda_c+\delta(\beta)+(4\beta-1)t\gamma)\right)\\ &\approx \lambda_c^{t+1}\left(1+\left(\beta-\frac{1}{4}\right)\sum_{\tau=0}^{t}\frac{4(\Delta+\tau\gamma)}{\lambda_c}\right)\\ &= \lambda_c^{t+1}\,(1+(4\beta-1)\,(t+1)(\Delta+\gamma t/2)\lambda_c)\end{aligned}$$

for all $t \in \{0, \ldots, T-1\}$, and

$$\begin{aligned}
\left(\bar{x} + \prod_{\tau=0}^{T-1} \hat{\lambda}_\tau\right)^{\frac{2\beta}{\kappa}} &= \exp\left(\frac{2\beta}{1-2\beta} \ln\left(\bar{x} + \exp \sum_{\tau=0}^{T-1} \ln(\lambda_c + (4\beta - 1)(\delta(\beta) + \tau\gamma))\right)\right) \\
&\approx \left(\bar{x} + \lambda_c^T\right)\left[1 + \left(\beta - \frac{1}{4}\right)\left(\frac{d}{d\beta}\left[\frac{2\beta}{1-2\beta}\right]_{\beta=\frac{1}{4}} \ln\left(\bar{x} + \lambda_c^T\right) + \frac{4\lambda_c^{T-1}\sum_{\tau=0}^{T-1}(\Delta + \tau\gamma)}{\bar{x} + \lambda_c^T}\right)\right] \\
&= \left(\bar{x} + \lambda_c^T\right)\left[1 + 2(4\beta - 1)\ln\left(\bar{x} + \lambda_c^T\right)\right] + (4\beta - 1) T \lambda_c^{T-1}(\Delta + (T-1)\gamma/2),
\end{aligned}$$

we get using (31)

$$u_T \approx \left(\frac{16\beta^2}{k\alpha}\right)^{\beta/\kappa} \frac{\bar{x} - \lambda_c^T\left(1 + (4\beta - 1)T(\Delta + \gamma(T-1)/2)\lambda_c\right)}{\left(\bar{x} + \lambda_c^T\right)\left[1 + 2(4\beta - 1)\ln\left(\bar{x} + \lambda_c^T\right)\right] + (4\beta - 1) T \lambda_c^{T-1}(\Delta + (T-1)\gamma/2)},$$

and a linear approximation of the last term gives

$$u_T \approx \left(\frac{16\beta^2}{k\alpha}\right)^{\beta/\kappa}\left[\frac{\bar{x} - \lambda_c^T}{\bar{x} + \lambda_c^T}\left[1 - 2(4\beta - 1)\ln\left(\bar{x} + \lambda_c^T\right)\right] - \bar{x}\lambda_c^{T-1} T(4\beta - 1)\frac{2\Delta + \gamma(T-1)}{(\bar{x} + \lambda_c^T)^2}\right].$$

Denote by $G_\beta(\lambda_0, T)$ the social planner objective and $\hat{G}_\beta(\Delta, T)$ its approximation when $\beta = 1/4 \pm \varepsilon$, $\varepsilon > 0$ small,

$$\begin{aligned}
\hat{G}_\beta(\Delta, T) = (1 - 2\beta)&\left(\frac{16\beta^2}{k\alpha}\right)^{\beta/\kappa} \\
&\times \left(T\left(\frac{\lambda_c - 1}{\lambda_c + 1}(1 - (4\beta - 1)\left[(T-1)\ln\lambda_c + 2\ln(\lambda_c + 1)\right]) + 2(4\beta - 1)\frac{\Delta + (T-1)\gamma}{(\lambda_c + 1)^2}\right)\right. \\
&\left. + \frac{\bar{x} - \lambda_c^T}{\bar{x} + \lambda_c^T}\left[1 - 2(4\beta - 1)\ln\left(\bar{x} + \lambda_c^T\right)\right] - \bar{x}\lambda_c^{T-1} T(4\beta - 1)\frac{2\Delta + \gamma(T-1)}{(\bar{x} + \lambda_c^T)^2}\right) - \nu T.
\end{aligned}$$

A Taylor expansion gives

$$\frac{\partial G_\beta(\lambda_0, T)}{\partial \lambda_0} \approx \frac{\partial G_{1/4}(\lambda_c, T)}{\partial \lambda_0} + (\lambda_0 - \lambda_c)\frac{\partial^2 G_{1/4}(\lambda_c, T)}{\partial \lambda_0^2} + (\beta - 1/4)\frac{\partial^2 G_{1/4}(\lambda_c, T)}{\partial \lambda_0 \partial \beta}$$

where the first term is null by definition of λ_c and $\lambda_0 \approx \lambda_c + \Delta(4\beta - 1)$. If ε is sufficiently small, the number of tiers T at the optimum is not affected, and the optimality condition $\partial G_\beta(\lambda_0, T)/\partial \lambda_0 = 0$ leads to

$$\Delta \approx -\frac{1}{4}\, \frac{\partial^2 G_{1/4}(\lambda_c, T)}{\partial \lambda_0 \partial \beta} \Bigg/ \frac{\partial^2 G_{1/4}(\lambda_c, T)}{\partial \lambda_0^2}$$

where the denominator is negative by concavity of G (at least locally at the optimum). The numerator can be approximated by

$$\frac{\partial^2 G_{1/4}(\lambda_c, T)}{\partial\lambda_0\partial\beta} \approx \frac{d}{d\beta}\left[\frac{\partial\hat{G}_\beta(0,T)}{\partial\Delta}\frac{d\Delta}{d\lambda_0}\right]_{\beta=1/4} = \frac{d}{d\beta}\left[\frac{\partial\hat{G}_\beta(0,T)}{\partial\Delta}\frac{1}{4\beta-1}\right]_{\beta=1/4}$$

where

$$\frac{\partial\hat{G}_\beta(\Delta,T)}{\partial\Delta}\frac{1}{4\beta-1} \approx (1-2\beta)\left(\frac{16\beta^2}{k\alpha}\right)^{\beta/\kappa} 2T\left(\frac{1}{(\lambda_c+1)^2} - \frac{\bar{x}\lambda_c^{T-1}}{(\bar{x}+\lambda_c^T)^2}\right).$$

Using $\bar{x} > \lambda_c^T$ we get

$$\begin{aligned}\frac{1}{(\lambda_c+1)^2} - \frac{\bar{x}\lambda_c^{T-1}}{(\bar{x}+\lambda_c^T)^2} < \frac{1}{(\lambda_c+1)^2} - \frac{\lambda_c^T\lambda_c^{T-1}}{(2\lambda_c^T)^2} &= \frac{1}{(\lambda_c+1)^2} - \frac{1}{4\lambda_c} \\ &= -\frac{(\lambda_c-1)^2}{4\lambda_c(\lambda_c+1)^2} < 0,\end{aligned}$$

and we have

$$\begin{aligned}\frac{d}{d\beta}\left[(1-2\beta)\left(\frac{16\beta^2}{k\alpha}\right)^{\beta/\kappa}\right] &= -2\left(\frac{16\beta^2}{k\alpha}\right)^{\beta/\kappa} + (1-2\beta)\left(\frac{16\beta^2}{k\alpha}\right)^{\beta/\kappa} \\ &\quad\times\left(\frac{d}{d\beta}\left[\frac{\beta}{1-2\beta}\right](4\ln 2 + 2\ln\beta - \ln\alpha k) + \frac{\beta}{1-2\beta}\frac{2}{\beta}\right) \\ &= \left(\frac{16\beta^2}{k\alpha}\right)^{\beta/\kappa}\frac{4\ln 2 + 2\ln\beta - \ln\alpha k}{1-2\beta},\end{aligned}$$

hence

$$\frac{d}{d\beta}\left[(1-2\beta)\left(\frac{16\beta^2}{k\alpha}\right)^{\beta/\kappa}\right]_{\beta=1/4} = \frac{-2\ln\alpha k}{\sqrt{k\alpha}}$$

which is positive if $\alpha k < 1$. Under this condition and assuming β close to 1/4, we have $\Delta < 0$ and consequently $\lambda_0 > \lambda_c$ iff $\beta < 1/4$.

C Proof of Proposition 1

Using $x_{-1} = 1$ and $x_t = \prod_{\tau=0}^t \lambda_{ct}$ for $t \in \{0, \ldots, T\}$, we get

$$\frac{x_{t+1} - x_t}{x_t - x_{t-1}} = \lambda_{ct}\frac{\lambda_{ct+1} - 1}{\lambda_{ct} - 1} = \left(1 + \frac{\lambda_{ct+1} - \lambda_{ct}}{\lambda_{ct} - 1}\right)\lambda_{ct}$$

for all $t \in \{0, \dots, T-1\}$. We thus have $x_{t+1} - x_t = \lambda_c(x_t - x_{t-1})$ when $\beta = 1/4$. For $\beta \approx 1/4$, we get using $\lambda_{ct} = \lambda_c + (4\beta - 1)(\Delta + t\gamma)$ and a first-order Taylor expansion that

$$\frac{\lambda_{ct+1} - \lambda_{ct}}{\lambda_{ct} - 1} \approx \frac{(4\beta - 1)\gamma}{\lambda_c + (4\beta - 1)(\Delta + t\gamma) - 1} \approx \frac{(4\beta - 1)\gamma}{\lambda_c - 1},$$

hence

$$\frac{x_{t+1} - x_t}{x_t - x_{t-1}} \approx \left(1 + \frac{(4\beta - 1)\gamma}{\lambda_c - 1}\right) \lambda_{ct}$$

which increases when $\beta > 1/4$ since $\boldsymbol{\lambda}_c$ increases, and the reverse otherwise. From (25), we get

$$\frac{n_{t+1}}{n_t} = \left(\lambda_t \frac{\lambda_{t+1} + 1}{\lambda_t + 1}\right)^{1/\kappa} = \lambda_t^{1/\kappa} \left(1 + \frac{\lambda_{t+1} - \lambda_t}{\lambda_t + 1}\right)^{1/\kappa},$$

for all $t \in \{0, \dots, T-2\}$, hence $n_{t+1}/n_t = \lambda_c^2$ when $\beta = 1/4$, and $n_{t+1}/n_t > \lambda_t^2$ with

$$\frac{n_{t+1}/n_t}{n_t/n_{t-1}} = \left(\frac{\lambda_t}{\lambda_{t-1}} \frac{\lambda_{t+1} + 1}{\lambda_{t-1} + 1}\right)^{1/\kappa} > 1$$

when $\beta > 1/4$ since the sequence $\{\lambda_t\}_{t=0}^{T-1}$ is increasing, and the reverse otherwise.

From (24), it comes

$$\begin{aligned}
\frac{g_t}{g_{t+1}} &= \frac{x_t - x_{t-1}}{x_{t+1} - x_t} \left(\frac{n_{t+1}}{n_t}\right)^2 = \frac{1}{\lambda_t} \frac{\lambda_t - 1}{\lambda_{t+1} - 1} \left(\lambda_t \frac{\lambda_{t+1} + 1}{\lambda_t + 1}\right)^{2/\kappa} \\
&= \lambda_t^{(1+2\beta)/\kappa} \left(1 + \frac{\lambda_{t+1} - \lambda_t}{\lambda_t + 1}\right)^{2/\kappa} \left(1 - \frac{\lambda_{t+1} - \lambda_t}{\lambda_{t+1} - 1}\right)
\end{aligned}$$

which gives $g_t/g_{t+1} = \lambda_c^3$ when $\beta = 1/4$. When $\beta \approx 1/4$, we have $g_t/g_{t+1} \approx \lambda_t^{(1+2\beta)/\kappa}$ and using

$$\begin{aligned}
\frac{\lambda_t - 1}{\lambda_{t+1} - 1} \frac{\lambda_{t+1} + 1}{\lambda_t + 1} &\approx \left(1 + \frac{(4\beta - 1)\gamma}{\lambda_c + (4\beta - 1)(\Delta + t\gamma) + 1}\right) \left(1 - \frac{(4\beta - 1)\gamma}{\lambda_c + (4\beta - 1)(\Delta + (t+1)\gamma) - 1}\right) \\
&\approx 1 - \frac{2(4\beta - 1)\gamma}{\lambda_c^2 - 1}
\end{aligned}$$

neglecting second-order terms, it comes

$$\frac{g_t}{g_{t+1}} \approx \left(1 - \frac{2(4\beta - 1)\gamma}{\lambda_c^2 - 1}\right) \left(\lambda_t \frac{\lambda_{t+1} + 1}{\lambda_t + 1}\right)^{(1+2\beta)/\kappa}.$$

We thus have

$$\frac{g_{t-1}/g_t}{g_t/g_{t+1}} \approx \left(\frac{\lambda_t}{\lambda_{t-1}}\frac{\lambda_{t+1}+1}{\lambda_{t-1}+1}\right)^{(1+2\beta)/\kappa},$$

implying that g_t decreases at an increasing rate when $\beta > 1/4$, and at a decreasing rate otherwise. For the per capita capacity, we have

$$\begin{aligned}\frac{n_t g_t}{n_{t+1}g_{t+1}} &= \frac{x_t - x_{t-1}}{x_{t+1}-x_t}\frac{n_{t+1}}{n_t} = \frac{1}{\lambda_t}\frac{\lambda_t - 1}{\lambda_{t+1}-1}\left(\lambda_t\frac{\lambda_{t+1}+1}{\lambda_t+1}\right)^{1/\kappa}\\ &= \lambda_t^{2\beta/\kappa}\left(1-\frac{\lambda_{t+1}-\lambda_t}{\lambda_{t+1}-1}\right)\left(1+\frac{\lambda_{t+1}-\lambda_t}{\lambda_t+1}\right)^{1/\kappa},\end{aligned}$$

hence $n_t g_t/n_{t+1}g_{t+1} = \lambda_c$ when $\beta = 1/4$. When $\beta \approx 1/4$, $n_t g_t/n_{t+1}g_{t+1} \approx \lambda_t^{2\beta/\kappa}$ and using

$$\begin{aligned}\frac{n_t g_t}{n_{t+1}g_{t+1}} &= \lambda_t^{2\beta/\kappa}\left(1-\frac{\lambda_{t+1}-\lambda_t}{\lambda_{t+1}-1}\right)\left(1+\frac{\lambda_{t+1}-\lambda_t}{\lambda_t+1}\right)^{1/\kappa}\\ &\approx \left(1-\frac{2(4\beta-1)\gamma}{\lambda_c^2-1}\right)\lambda_t^{2\beta/\kappa}\left(1+\frac{\lambda_{t+1}-\lambda_t}{\lambda_t+1}\right)^{2\beta/\kappa}\end{aligned}$$

neglecting second-order terms, $n_t g_t$ decreases at an increasing rate when $\beta > 1/4$ and at a decreasing rate when $\beta < 1/4$. Finally, it comes from (30)

$$\frac{u_{t+1}}{u_t} = \left(1+\frac{\lambda_{t+1}-\lambda_t}{\lambda_t-1}\right)\left(1-\frac{\lambda_{t+1}-\lambda_t}{\lambda_{t+1}+1}\right)^{2\beta/\kappa}\lambda_t^{(1-4\beta)/\kappa}$$

which gives $u_{t+1} = u_t$ when $\beta = 1/4$. When $\beta \approx 1/4$, $u_{t+1}/u_{t+1} \approx \lambda_t^{(1-4\beta)/\kappa}$, and as $\lambda_t > 1$, $u_{t+1} > u_t$ when $\beta < 1/4$ and the reverse otherwise.

D Proof of Proposition 2

Denoting $W_t \equiv W(x_{t-1}, x_t, g_t, n_t)$ and differentiating (5) w.r.t. g_t gives

$$\frac{\partial W_t}{\partial g_t} = \beta\frac{u_t}{g_t} - \frac{n_t k\left(x_t + x_{t-1}\right)}{2} \tag{35}$$

which is null at the social planner's optimum. Hence, if assigned scope $(x_{t-1}, x_t]$, the jurisdiction would chose the same administration's capacity as that of the social planner. Differentiating (5) w.r.t. x_{t-1} and x_t gives

$$\frac{\partial W_t}{\partial x_{t-1}} = -\frac{(1-\beta)u_t}{x_t - x_{t-1}} + \frac{\alpha x_{t-1}}{4n_t} - \frac{n_t k g_t}{2}, t \in \{1, \ldots, T\} \tag{36}$$

and

$$\frac{\partial W_t}{\partial x_t} = \frac{(1-\beta)u_t}{x_t - x_{t-1}} - \frac{\alpha x_t}{4n_t} - \frac{n_t k g_t}{2}, t \in \{0, \ldots, T-1\}. \tag{37}$$

Using (23) to substitute for the first term of (36) evaluated at the social planner's optimum, we get

$$\begin{aligned}
\frac{\partial W_t}{\partial x_{t-1}} &= \frac{\alpha x_{t-1}}{4n_t} - \frac{n_t k g_t}{2}\left(\frac{1-\beta}{\beta}\frac{x_t + x_{t-1}}{x_t - x_{t-1}} + 1\right) \\
&\leq \frac{\alpha x_{t-1}}{4n_t} - \frac{n_t k g_t}{2}\left(\frac{x_t + x_{t-1}}{x_t - x_{t-1}} + 1\right) = \frac{\alpha x_{t-1}}{4n_t} - \frac{n_t k g_t x_t}{x_t - x_{t-1}} \\
&= \frac{n_t x_t}{x_t - x_{t-1}}\left(\frac{x_{t-1}}{x_t}\frac{\alpha(x_t - x_{t-1})}{4n_t^2} - k g_t\right) \\
&< \frac{n_t x_t}{x_t - x_{t-1}}\left(\frac{\alpha(x_t - x_{t-1})}{4n_t^2} - k g_t\right) \\
&= 0
\end{aligned}$$

where the first inequality comes from $\beta \leq 1/2$, the second one from $x_{t-1} < x_t$ and the last equality from (19).

Using (23) to substitute for the first term of (37) evaluated at the social planner optimum give

$$\begin{aligned}
\frac{\partial W_t}{\partial x_t} &= \frac{n_t k g_t}{2}\left(\frac{1-\beta}{\beta}\frac{x_t + x_{t-1}}{x_t - x_{t-1}} - 1\right) - \frac{\alpha x_t}{4n_t} \\
&= \frac{n_t k g_t}{2}\left(\frac{1-\beta}{\beta}\frac{x_t + x_{t-1}}{x_t - x_{t-1}} - 1\right) - \frac{n_t k g_t x_t}{x_t - x_{t-1}} \\
&= \frac{n_t k g_t}{2}\left(\frac{1-\beta}{\beta}\frac{x_t + x_{t-1}}{x_t - x_{t-1}} - \frac{3x_t - x_{t-1}}{x_t - x_{t-1}}\right) \\
&= \frac{n_t k g_t}{2\beta}\frac{x_t(1-4\beta) + x_{t-1}}{x_t - x_{t-1}} \\
&= \frac{n_t k g_t}{2\beta}\frac{x_{t-1}}{x_t - x_{t-1}}[\lambda_t(1-4\beta) + 1]
\end{aligned}$$

which is positive for all $\beta \leq 1/4$ and more generally if $\lambda_t(1-4\beta)+1 > 0$. As the sequence $\{\lambda_{ct}\}_{t=0}^{T}$ is increasing when $\beta > 1/4$, it suffices that this inequality is satisfied for λ_{c0}, hence $\beta < (1/\lambda_{c0} + 1)/4$.

E Proof of Lemma 3

Using (20) to substitute for u_t in (37) and equalizing to zero gives

$$0 = \frac{\partial W_t}{\partial x_t} = \frac{1-\beta}{\beta} \frac{n_t k g_t (x_t + x_{t-1})}{2(x_t - x_{t-1})} - \frac{\alpha x_t}{4 n_t} - \frac{n_t k g_t}{2},$$

hence

$$g_t = \frac{\alpha \beta x_t (x_t - x_{t-1})}{2 k n_t^2 (\kappa x_t + x_{t-1})}, t \in \{0, \dots, T-1\}, \tag{38}$$

Identifying (38) with (22), it comes

$$\begin{aligned} n_t &= \left[\frac{\alpha^{1-\beta} k^{\beta}}{2^{2-\beta} \beta^{\beta}} \left(\frac{x_t}{\kappa x_t + x_{t-1}} \right)^{1-\beta} (x_t + x_{t-1}) \right]^{1/\kappa}, \\ &= \left(\frac{2\beta x_t}{\kappa x_t + x_{t-1}} \right)^{(1-\beta)/\kappa} K \left(x_t + x_{t-1} \right)^{1/\kappa}, \end{aligned} \tag{39}$$

for $t \in \{0, \dots, T-1\}$. As (22) and (21) are derived from (20), they both hold in that case. We have to solve

$$\max_{T, \{n_t, x_t, g_t\}_{t=0}^{T}} \left\{ \sum_{t=0}^{T} W_t - \nu T : x_{-1} = 1, x_T = \bar{x}, (20), (38) \right\}.$$

Neglecting the boundary conditions on x_{-1} and x_T, and denoting by χ_t and θ_t the multipliers corresponding to the constraints (20) and (38) respectively, the Lagrangian of this program is

$$L = \sum_{t=0}^{T} \left[W_t + \chi_t \left(\beta u_t - \frac{n_t k g_t (x_t + x_{t-1})}{2} \right) + \theta_t \left(g_t - \frac{\alpha \beta x_t (x_t - x_{t-1})}{2 k n_t^2 (\kappa x_t + x_{t-1})} \right) \right] - \nu T.$$

Using $\partial W_t / \partial g_t = 0$ and (20), the FOCs w.r.t. g_t and n_t give

$$\theta_t = \chi_t \left(\frac{n_t k (x_t + x_{t-1})}{2} - \beta^2 \frac{u_t}{g_t} \right) = \chi_t (1-\beta) \frac{n_t k (x_t + x_{t-1})}{2}, \tag{40}$$

and

$$\frac{\partial W_t}{\partial n_t} + \theta_t \frac{2 g_t}{n_t} - \chi_t \frac{g_t k (x_t + x_{t-1})}{2} = 0, \tag{41}$$

for $t \in \{0, \ldots, T-1\}$, respectively. Using (40), the latter simplifies to

$$\frac{\partial W_t}{\partial n_t} = -\chi_t \kappa \frac{g_t k\left(x_t + x_{t-1}\right)}{2}, \tag{42}$$

where

$$\frac{\partial W_t}{\partial n_t} = \frac{\alpha\left(x_t^2 - x_{t-1}^2\right)}{8n_t^2} - \frac{kg_t\left(x_t + x_{t-1}\right)}{2}. \tag{43}$$

We thus have

$$\frac{\alpha\left(x_t^2 - x_{t-1}^2\right)}{8n_t^2} = (1 - \chi_t\kappa)\frac{g_t k\left(x_t + x_{t-1}\right)}{2}$$

and using (38), we arrive at

$$\chi_t = \frac{(2\beta - \kappa)x_t - x_{t-1}}{2\beta\kappa_1 x_t} \tag{44}$$

for $t \in \{0, \ldots, T-1\}$. Hence, compared to (23), we get using (20),

$$\frac{\alpha\left(x_t^2 - x_{t-1}^2\right)}{8n_t} = \frac{\kappa x_t + x_{t-1}}{2\beta x_t}\frac{n_t g_t k\left(x_t + x_{t-1}\right)}{2} = \frac{\kappa x_t + x_{t-1}}{x_t}\frac{u_t}{2} \tag{45}$$

for all $t \in \{0, \ldots, T-1\}$ under decentralization. The average access cost is thus greater than the provision cost iff $\beta < (x_{t-1}/x_t + 1)/4$.

Using $\partial W_t/\partial x_t = 0$, the FOC w.r.t. x_t simplifies to

$$\begin{aligned} 0 = &\frac{\partial W_{t+1}}{\partial x_t} - \theta_t\frac{dg_t}{dx_t} + \chi_t\left(\frac{(1-\beta)\beta u_t}{x_t - x_{t-1}} - \frac{n_t k g_t}{2}\right) \\ &- \theta_{t+1}\frac{dg_{t+1}}{dx_t} + \chi_{t+1}\left(-\frac{(1-\beta)\beta u_{t+1}}{x_{t+1} - x_t} - \frac{n_{t+1}kg_{t+1}}{2}\right), \end{aligned} \tag{46}$$

for $t \in \{0, \ldots, T-1\}$, where

$$\begin{aligned} \frac{dg_t}{dx_t} &= \frac{d}{dx_t}\left[\frac{\alpha\beta x_t(x_t - x_{t-1})}{2kn_t^2(\kappa x_t + x_{t-1})}\right] = \frac{g_t[x_t(\kappa x_t + 2x_{t-1}) - x_{t-1}^2]}{x_t(x_t - x_{t-1})(\kappa x_t + x_{t-1})}, \\ \frac{dg_{t+1}}{dx_t} &= \frac{d}{dx_t}\left[\frac{\alpha\beta x_{t+1}(x_{t+1} - x_t)}{2kn_{t+1}^2(\kappa x_{t+1} + x_t)}\right] = \frac{-2(1-\beta)x_{t+1}g_{t+1}}{(\kappa x_{t+1} + x_t)(x_{t+1} - x_t)}, \end{aligned}$$

and, using (20),

$$\frac{(1-\beta)\beta u_t}{x_t - x_{t-1}} - \frac{n_t k g_t}{2} = \frac{n_t k g_t}{2}\frac{(2-\beta)x_{t-1} - \beta x_t}{x_t - x_{t-1}},$$

and
$$\frac{(1-\beta)\beta u_{t+1}}{x_{t+1}-x_t}+\frac{n_{t+1}kg_{t+1}}{2}=\frac{n_{t+1}kg_{t+1}}{2}\frac{(2-\beta)x_{t+1}-\beta x_t}{x_{t+1}-x_t}.$$

Finally, using $\partial W_t/\partial x_t = 0$, we get from (36) and (37) that

$$\frac{\partial W_t}{\partial x_{t-1}}=-\frac{\alpha}{4n_t}(x_t-x_{t-1})-kn_tg_t. \tag{47}$$

Using (38) and (44), we obtain

$$\frac{\partial W_{t+1}}{\partial x_t}=-kn_{t+1}g_{t+1}\chi_{t+1}\frac{\kappa(x_{t+1}+x_t)}{(2\beta-\kappa)x_{t+1}-x_t}.$$

We can thus rewrite (46) as

$$0=-\chi_t n_t g_t\left((1-\beta)\frac{(x_t+x_{t-1})}{2}\frac{x_t(\kappa x_t+2x_{t-1})-x_{t-1}^2}{x_t(x_t-x_{t-1})(\kappa x_t+x_{t-1})}-\frac{(2-\beta)x_{t-1}-\beta x_t}{2(x_t-x_{t-1})}\right)$$
$$-n_{t+1}g_{t+1}\chi_{t+1}\left(-\frac{(1-\beta)^2\left(x_t+x_{t+1}\right)x_{t+1}}{(\kappa x_{t+1}+x_t)(x_{t+1}-x_t)}+\frac{(2-\beta)x_{t+1}-\beta x_t}{2(x_{t+1}-x_t)}+\frac{\kappa(x_{t+1}+x_t)}{(2\beta-\kappa)x_{t+1}-x_t}\right)$$

for $t\in\{0,\ldots,T-1\}$. Using $x_t=\lambda_t x_{t-1}=\prod_{\tau=0}^{t}\lambda_\tau$ yields

$$\chi_t=\frac{(2\beta-\kappa)x_t-x_{t-1}}{2\beta\kappa_1 x_t}=\frac{(4\beta-1)\lambda_t-1}{2\beta\kappa_1\lambda_t},$$

and from (38),

$$n_tg_t=\frac{\alpha\beta(\lambda_t-1)\prod_{\tau=0}^{t}\lambda_\tau}{2k[\kappa\lambda_t+1]n_t},$$

which gives

$$\frac{n_{t+1}g_{t+1}}{n_tg_t}=\frac{\frac{\lambda_{t+1}-1}{\kappa\lambda_{t+1}+1}}{\frac{\lambda_t-1}{\kappa\lambda_t+1}}\lambda_{t+1}\frac{n_t}{n_{t+1}}.$$

Replacing and collecting terms, (46) becomes

$$0=-\chi_t\left((1-\beta)\frac{(\lambda_t+1)}{2}\frac{\lambda_t(\kappa\lambda_t+2)-1}{\lambda_t(\lambda_t-1)(\kappa\lambda_t+1)}-\frac{(2-\beta)-\beta\lambda_t}{2(\lambda_t-1)}\right)$$
$$-\frac{\frac{\lambda_{t+1}-1}{\kappa\lambda_{t+1}+1}}{\frac{\lambda_t-1}{\kappa\lambda_t+1}}\lambda_{t+1}\frac{n_t}{n_{t+1}}\chi_{t+1}\left(-\frac{(1-\beta)^2\left(1+\lambda_{t+1}\right)\lambda_{t+1}}{(\kappa\lambda_{t+1}+1)(\lambda_{t+1}-1)}+\frac{(2-\beta)\lambda_{t+1}-\beta}{2(\lambda_{t+1}-1)}+\frac{\kappa(\lambda_{t+1}+1)}{(2\beta-\kappa)\lambda_{t+1}-1}\right),$$

which yields, using (44),

$$\frac{n_{t+1}}{n_t} = -\frac{\frac{\lambda_{t+1}-1}{\kappa\lambda_{t+1}+1}}{\frac{\lambda_t-1}{\kappa\lambda_t+1}} \frac{1+\lambda_{t+1}+\frac{\beta[(4\beta-1)\lambda_{t+1}-1](1-\lambda_{t+1})}{2\kappa_1(\kappa\lambda_{t+1}+1)}}{\frac{[(4\beta-1)\lambda_t-1][\lambda_t(\kappa\lambda_t+1)+(1-\beta)(\lambda_t+1)]}{2\kappa_1\lambda_t^2(\kappa\lambda_t+1)}}.$$

Using (39) gives

$$\frac{\left[(\lambda_{t+1}+1)\left(\frac{\lambda_{t+1}}{\kappa\lambda_{t+1}+1}\right)^{1-\beta}\right]^{1/\kappa}}{\left[(\lambda_t+1)\left(\frac{\lambda_t}{\kappa\lambda_t+1}\right)^{1-\beta}\right]^{1/\kappa}}\lambda_t^{1/\kappa} = -\frac{\frac{\lambda_{t+1}-1}{\kappa\lambda_{t+1}+1}}{\frac{\lambda_t-1}{\kappa\lambda_t+1}} \frac{1+\lambda_{t+1}+\frac{\beta[(4\beta-1)\lambda_{t+1}-1](1-\lambda_{t+1})}{2\kappa_1(\kappa\lambda_{t+1}+1)}}{\frac{[(4\beta-1)\lambda_t-1][\lambda_t(\kappa\lambda_t+1)+(1-\beta)(\lambda_t+1)]}{2\kappa_1\lambda_t^2(\kappa\lambda_t+1)}} \quad (48)$$

for $t \in \{0, \ldots, T-2\}$, that defines implicitly a first-order recurrence equation $\lambda_{t+1} = F_d(\lambda_t) = F_d^t(\lambda_0)$, hence the sub-sequence $\{\lambda_t\}_{t=1}^{T-1}$ given λ_0: $\lambda_t = F_d^t(\lambda_0) \equiv F_d(F_d^{t-1}(\lambda_0))$ (with $F_d^0 = Id$). The last term of the sequence is deduced from the boundary condition $x_T = \bar{x}$ which gives $\lambda_T = \bar{x}/\prod_{t=0}^{T-1} F_d^t(\lambda_0)$. Using (39) gives

$$\frac{n_{t+1}}{n_t} = \left(1 - \frac{\kappa(\lambda_{t+1}-\lambda_t)}{\kappa\lambda_{t+1}+1}\right)^{(1-\beta)/\kappa} \left(1 + \frac{\lambda_{t+1}-\lambda_t}{\lambda_t+1}\right)^{1/\kappa} \left(\lambda_{t+1}^{1-\beta}\lambda_t^{\beta}\right)^{1/\kappa}.$$

Plugging (39) into (21) and (22) yields

$$\begin{aligned} u_t &= \left(\frac{8\beta}{\alpha k}\frac{\kappa x_t + x_{t-1}}{x_t\left(x_t+x_{t-1}\right)^2}\right)^{\beta/\kappa}(x_t - x_{t-1}) \\ &= \left(\frac{8\beta}{\alpha k}\frac{\kappa\lambda_t+1}{\lambda_t\left(\lambda_t+1\right)^2}\right)^{\beta/\kappa}(\lambda_t-1)\prod_{\tau=0}^{t-1}\lambda_t^{(1-4\beta)/\kappa}, \end{aligned} \quad (49)$$

and

$$g_t = \left(\frac{8\beta}{\alpha k}\frac{\kappa x_t + x_{t-1}}{x_t\left(x_t+x_{t-1}\right)^2}\right)^{1/\kappa}(x_t - x_{t-1}),$$

which give

$$\begin{aligned} \frac{u_{t+1}}{u_t} &= \left(\frac{\kappa\lambda_{t+1}+1}{\kappa\lambda_t+1}\right)^{\beta/\kappa}\left(\frac{\lambda_t}{\lambda_{t+1}}\right)^{\beta/\kappa}\frac{\lambda_{t+1}-1}{\lambda_t-1}\left(\frac{\lambda_t+1}{\lambda_{t+1}+1}\right)^{2\beta/\kappa}\lambda_t^{(1-4\beta)/\kappa} \\ &= \left(1+\frac{\kappa(\lambda_{t+1}-\lambda_t)}{\kappa\lambda_t+1}\right)^{\beta/\kappa}\left(\frac{\lambda_t}{\lambda_{t+1}}\right)^{\beta/\kappa}\left(1+\frac{\lambda_{t+1}-\lambda_t}{\lambda_t-1}\right)\left(1-\frac{\lambda_{t+1}-\lambda_t}{\lambda_{t+1}+1}\right)^{2\beta/\kappa}\lambda_t^{(1-4\beta)/\kappa} \end{aligned}$$

and

$$\frac{g_t}{g_{t+1}} = \left(\lambda_{t+1}\lambda_t^{2\beta}\right)^{1/\kappa}\left(1-\frac{\kappa(\lambda_{t+1}-\lambda_t)}{\kappa\lambda_{t+1}+1}\right)^{1/\kappa}\left(1-\frac{\lambda_{t+1}-\lambda_t}{\lambda_{t+1}-1}\right)\left(1+\frac{\lambda_{t+1}-\lambda_t}{\lambda_t+1}\right)^{2/\kappa}.$$

It comes from (45) that

$$W_t = \left(1-\beta-\frac{\kappa x_t + x_{t-1}}{2x_t}\right)u_t \tag{50}$$

for $t \in \{0, \ldots, T-1\}$. For tier T, as $\partial W_T/\partial n_T = 0$, it comes from (25) and (23) that $n_T = (\bar{x} + x_{T-1})^{1/\kappa} K$ and $W_T = \kappa u_T$, where u_T is given by (31). T satisfies

$$\prod_{t=0}^{T} F_d^t(\lambda_0) \geq \bar{x} > \prod_{t=0}^{T-1} F_d^t(\lambda_0), \tag{51}$$

and λ_0 and T are derived from solving

$$\max_{\lambda_0, T}\left\{\sum_{t=0}^{T-1}\left(1-\beta-\frac{\kappa F_d^t(\lambda_0)+1}{2F_d^t(\lambda_0)}\right)u_t + \kappa u_T - \nu T : (31),\ (49),\ (51)\right\}$$

where $F_d(\cdot)$ is implicitly defined by (48). With $\beta = 1/4$, (48) simplifies to $f_d(\lambda_{t+1}) = f_d(\lambda_t)$ where

$$f_d(\lambda) \equiv \frac{(\lambda+1)^2(\lambda+2)^{1/2}\lambda^{3/2}}{2(\lambda^2-1)(\lambda+2)+(\lambda-1)^2}.$$

One solution is thus given by $\lambda_t = \lambda_d$ for all $t = 0, \ldots, T-1$ (as f_d is convex, with a minimum at $\lambda \approx 2.12$, cyclical solutions are also possible). Using (49) and (50), we get

$$u_t = \left(\frac{\lambda_d+2}{\lambda_d}\right)^{1/2}\frac{1}{\sqrt{\alpha k}}\frac{\lambda_d-1}{\lambda_d+1} = \left(\frac{\lambda_d+2}{\lambda_d}\right)^{1/2} u_c(\lambda_d),$$

and

$$\begin{aligned} W_t &= \left(\frac{3}{4}-\frac{\lambda_d+2}{4\lambda_d}\right)\left(\frac{\lambda_d+2}{\lambda_d}\right)^{1/2} u_c(\lambda_d) \\ &= \left(\frac{\lambda_d-1}{\lambda_d}\right)\left(\frac{\lambda_d+2}{\lambda_d}\right)^{1/2}\frac{u_c(\lambda_d)}{2} \end{aligned}$$

for all $t \in \{0, \dots, T-1\}$. λ_d and T are derived by solving

$$\max_{\lambda_d, T} \left\{ T\Phi_d(\lambda_d)\frac{\lambda_d - 1}{\lambda_d + 1} + \frac{\bar{x} - \lambda_d^T}{\bar{x} + \lambda_d^T} - \nu T : \lambda_d^{T+1} \geq \bar{x} > \lambda_d^T \right\}$$

where $\Phi_d(\lambda) \equiv (\lambda - 1)(\lambda + 2)^{1/2}/\lambda^{3/2}$. Neglecting the constraints, the FOCs with respect to T and λ_d give

$$\Phi_d(\lambda_d)\frac{\lambda_d - 1}{\lambda_d + 1} - \nu - \frac{2\bar{x}\lambda_d^T \ln \lambda_d}{(\bar{x} + \lambda_d^T)^2} = 0$$

and

$$\Phi_d'(\lambda_d)\frac{\lambda_d - 1}{\lambda_d + 1} + \frac{2\Phi_d(\lambda_d)}{(\lambda_d + 1)^2} - \frac{2\bar{x}\lambda_d^{T-1}}{(\bar{x} + \lambda_d^T)^2} = 0,$$

respectively. Multiplying the last equation by $\lambda_d \ln \lambda_d$ and identifying with the first one yields

$$\lambda_d \ln \lambda_d \left(\Phi_d'(\lambda_d)\frac{\lambda_d - 1}{\lambda_d + 1} + \frac{2\Phi_d(\lambda_d)}{(\lambda_d + 1)^2} \right) = \frac{2\bar{x}\lambda_d^T \ln \lambda_d}{(\bar{x} + \lambda_d^T)^2} = \Phi_d(\lambda_d)\frac{\lambda_d - 1}{\lambda_d + 1} - \nu,$$

hence $g_d(\lambda_d) = \nu$ where

$$\begin{aligned} g_d(\lambda) &= [(\lambda^2 - 1)[\Phi_d(\lambda) - \Phi_d'(\lambda)\lambda \ln \lambda] - 2\Phi_d(\lambda)\lambda \ln \lambda]/(\lambda + 1)^2 \\ &= (\lambda - 1)[(\lambda^2 - 1)(\lambda + 2) - (\lambda + 3)(2\lambda + 1) \ln \lambda]/[\lambda^{3/2}(\lambda + 1)^2(\lambda + 2)^{1/2}]. \end{aligned}$$

As $g_d(1) = 0$ and using $\ln \lambda = -\ln(1 - (\lambda - 1)/\lambda) \geq (\lambda - 1)/\lambda$,

$$\begin{aligned} g_d(\lambda) &\leq (\lambda - 1)^2 [\lambda(\lambda + 1)(\lambda + 2) - (\lambda + 3)(2\lambda + 1)]/[\lambda^{5/2}(\lambda + 1)^2(\lambda + 2)^{1/2}] \\ &= (\lambda - 1)^2 [\lambda(\lambda + 1)(\lambda + 2) - (\lambda + 2)(2\lambda + 1) - (2\lambda + 1)]/[\lambda^{5/2}(\lambda + 1)^2(\lambda + 2)^{1/2}] \\ &< (\lambda - 1)^2 (\lambda + 2)^{1/2}[\lambda(\lambda + 1) - (2\lambda + 1) - 1]/[\lambda^{5/2}(\lambda + 1)^2] \\ &= (\lambda - 1)^2 (\lambda + 2)^{1/2} (\lambda - 2)(\lambda + 1)/[\lambda^{5/2}(\lambda + 1)^2], \end{aligned}$$

where the second inequality comes from $2\lambda + 1 > \lambda + 2$ when $\lambda > 1$. We thus have $g_d(\lambda) < 0$ for all $\lambda \in (1, 2]$. Also, as $g_d'(\lambda) = h(\lambda) \ln \lambda/[\lambda^{5/2}(\lambda + 1)(\lambda + 2)^{1/2}]$ where

$h(\lambda) = 2\lambda^5 + 10\lambda^4 - 3\lambda^3 - 41\lambda^2 - 31\lambda - 9$, $h(2) < 0$ and

$$\begin{aligned} h'(\lambda) &= 10\lambda^4 + 40\lambda^3 - 9\lambda^2 - 82\lambda - 31 \\ &> 40\lambda^3 - 9\lambda^2 - 82\lambda - 31 \\ &= (\lambda - 2)(40\lambda^2 + 71\lambda + 60) + 89 \end{aligned}$$

which is positive for all $\lambda \geq 2$, $g_d(\lambda)$ is increasing for all $\lambda \in \{\lambda > 2 : g_d(\lambda) \geq 0\}$. Hence, provided that $g_d(\bar{x}) > \nu, \lambda_d = g_d^{-1}(\nu)$ exits and is unique, greater than 2, and increases with ν. From (39) and $\lambda_d > 2$, we get

$$n_0 = \left(\frac{\lambda_d}{\lambda_d + 2}\right)^{3/2} \frac{\sqrt{k\alpha^3}}{2} (\lambda_d + 1)^2 > \left(\frac{1}{2}\right)^{3/2} \frac{\sqrt{k\alpha^3}}{2} (3)^2 = 9\sqrt{2^5 k\alpha^3}.$$

Hence, we must have $k\alpha^3 < 1/(81 \times 2^5) = 1/2592 \approx 4.10^{-4}$ to have $n_0 \leq 1$, and a sufficient condition for $n_0 \geq 2$ is $k\alpha^3 \geq 1/648 \approx 15.10^{-4}$.

F Proof of proposition 3

The generic AG problem is given by

$$\max_{\lambda, T} \left\{ T\Phi_j(\lambda)\frac{\lambda - 1}{\lambda + 1} + \frac{\bar{x} - \lambda^T}{\bar{x} + \lambda^T} - \nu T : \lambda^{T+1} \geq \bar{x} > \lambda^T \right\},$$

$j \in \{c, d\}$, with $\Phi_c(\lambda) = 1$ and $\Phi_d(\lambda) = (\lambda+2)^{1/2}(\lambda - 1)/\lambda^{3/2}$. As shown in Lemmas 2 and 3, the solution is given by $\lambda_j = g_j^{-1}(\nu)$ where $g_j(\lambda) = [(\lambda^2-1)[\Phi_j(\lambda) - \Phi_j'(\lambda)\lambda \ln \lambda] - 2\Phi_j(\lambda)\lambda \ln \lambda]/(\lambda + 1)^2$ is increasing for all $\lambda \in \{\lambda > 1 : g_j(\lambda) \geq 0\}$. We thus have $\lambda_j > \lambda_m$ iff $g_m(\lambda_j) > g_m(\lambda_m) = \nu = g_j(\lambda_j)$, hence $g_m(\lambda_j) - g_j(\lambda_j) > 0$. This is true for all $\nu \in (0, 1)$ if $g_m(\lambda) - g_j(\lambda) > 0$ for all $\lambda > 1$, i.e.

$$h(\lambda)(\Phi_m(\lambda) - \Phi_j(\lambda)) - (\Phi_m'(\lambda) - \Phi_j'(\lambda))(\lambda^2 - 1)\lambda \ln \lambda > 0$$

for all $\lambda > 1$, where $h(\lambda) = \lambda^2 - 1 - 2\lambda \ln \lambda$. As $h'(\lambda) = 2(\lambda - \ln \lambda - 1)$ and $h''(\lambda) = 2(1 - 1/\lambda) > 0$ when $\lambda > 1$, we have $h(1) = 0, h'(1) = 0$ and $h'(\lambda) > 0$ for $\lambda > 1$, implying that $h(\lambda) > 0$ for $\lambda > 1$. Comparing decentralization to first-best ($m = c$ and $j = d$), as $(\lambda+2)(\lambda - 1)^2 = \lambda^3 - 3\lambda + 2 < \lambda^3$ when $\lambda > 2/3$, we have $\Phi_d(\lambda) < 1 = \Phi_c(\lambda)$ for all $\lambda > 1$. As $\Phi_c'(\lambda) = 0 < \Phi_d'(\lambda) = 3/[\lambda^{5/2}(\lambda + 2)^{1/2}]$, we thus have $\lambda_d > \lambda_c$.

Table 1: Simulations of AGs

	First-Best					Decentralization				
	$\lambda_c = 1.94$, $\sum_t W_{ct} = 1, \sum_t \hat{W}_{ct} = .93$					$\lambda_d = 4.07$, $\sum_t W_{dt} = .8, \sum_t \hat{W}_{dt} = .76$				
t	x_{ct}	n_{ct}	$\hat{n}_{ct}$	g_{ct}	$\hat{W}_{ct}$	x_{dt}	n_{dt}	$\hat{n}_{dt}$	g_{dt}	$\hat{W}_{dt}$
0	1.94	1.21	1	$1.18 \cdot 10^{-2}$	0.14	4.07	1.97	2	$6.61 \cdot 10^{-3}$	0.25
1	3.77	4.56	5	$9.15 \cdot 10^{-4}$	0.14	16.57	32.7	33	$9.80 \cdot 10^{-5}$	0.25
2	7.33	17.21	17	$1.54 \cdot 10^{-4}$	0.14	67.46	541.87	542	$1.45 \cdot 10^{-6}$	0.25
3	14.23	64.92	65	$2.04 \cdot 10^{-5}$	0.14	100	1,088.45	1,088	$1.46 \cdot 10^{-7}$	$4.33 \cdot 10^{-2}$
4	27.63	244.87	245	$2.79 \cdot 10^{-6}$	0.14	—	—	—	—	—
5	53.66	923.61	924	$3.81 \cdot 10^{-7}$	0.14	—	—	—	—	—
6	100	3,299.94	3,300	$5.32 \cdot 10^{-8}$	0.13	—	—	—	—	—

Parameter values: $\alpha = .25, k = 5, \nu = .01, \bar{x} = 100$.

Table 2: Summary statistics

Variables	# of Obs.	Mean	SD	Min	Max
ln of expenditure ratio (Y)[1]	1950	4.93	0.95	2.23	6.79
4 $\times$ ln of population ratio (P)[2]	1950	17.58	4.03	8.43	25.43
Income per capita[2]	1950	25.41	12.58	5.35	68.33
Pop. density[2,3]	1950	148.33	190.21	0.6	1024.4
Dem. majority chambers[4]	1862*	0.49	0.5	0	1
Rep. majority chambers[4]	1862*	0.3	0.46	0	1

*As Nebraska is a non-partisan, unicameral legislature, it is excluded from the regressions that include partisan composition.

Sources: [1]the Tax Policy Center & the U.S. Government Publishing Office, [2]the Bureau of Economic Analysis, [3]the Census Bureau, [4]the National Conference of State Legislatures.

Table 3: Estimation results

Model specification	(1)	(2)	(3)	(4)
All states ($\hat{\beta}$)	0.184***	0.182***		
	(0.0238)	(0.0254)		
Midwest ($\hat{\beta}_1$)			0.279***	
			(0.0624)	
Northeast ($\hat{\beta}_2$)			0.182**	
			(0.0579)	
South ($\hat{\beta}_3$)			0.238***	
			(0.0309)	
West ($\hat{\beta}_4$)			0.147***	
			(0.0255)	
Northeast & West ($\hat{\beta}_A$)				0.153***
				(0.0235)
Midwest & South ($\hat{\beta}_B$)				0.251***
				(0.0342)
Income per capita	-0.0025	-0.0029	-0.0026	-0.0021
	(0.0027)	(0.0029)	(0.0031)	(0.0029)
Pop. density	-0.0006	-0.0006	-0.00006	-0.00004
	(0.0005)	(0.0005)	(0.0006)	(0.0006)
Dem. chambers		0.0224	0.0185	0.0197
		(0.0136)	(0.0138)	(0.0132)
Rep. chambers		0.0136	0.0178	0.0194
		(0.0166)	(0.0168)	(0.0161)
Observations	1950	1862	1862	1862
Hausman's test	23.82***	36.36***	50.33***	65.22***
F-Stat	473.2***	479.8***	845.2***	1021.7***

Notes:*** indicates statistical significance at 1%, ** at 5%, and * at 10% level. Standard errors clustered at the state level are in parentheses. All estimations include year fixed effects. Estimates of the constant term and of the time-dummies are omitted.

Table 4: Partisan views of government

Partisanship	Federal	State	Local
Rep.	13%	57%	63%
Ind.	27%	59%	60%
Dem.	41%	56%	67%
Overall	28%	57%	63%

Source: Pew Research Center (2013).
Answers to the question "Would you say your overall opinion of (government level) is very favorable, mostly favorable, mostly unfavorable, or very unfavorable?" Reported are the proportions corresponding to the sum of "very favorable" and "mostly favorable" by partisanship. Survey conducted March 13-17, 2013 among 1,501 adults living in all 50 U.S. states and the District of Columbia.

Table 5: Equality tests

H_0	$\hat{\beta}_1 = \hat{\beta}_2$	$\hat{\beta}_1 = \hat{\beta}_3$	$\hat{\beta}_1 = \hat{\beta}_4$	$\hat{\beta}_2 = \hat{\beta}_3$	$\hat{\beta}_2 = \hat{\beta}_4$	$\hat{\beta}_3 = \hat{\beta}_4$	$\hat{\beta}_A = \hat{\beta}_B$
F-test	3.07	0.52	3.28	0.88	0.27	4.76	5.89
p-value	0.0862	0.4726	0.0764	0.3517	0.6041	0.0341	0.0190