Big-City Governments

by

Brendan O'Flaherty, Columbia University

May 1995

Discussion Paper Series No. 731
Big-City Governments

Brendan O'Flaherty  
Department of Economics  
Columbia University  
New York, New York 10027  
Second draft--May 15, 1995

I have benefitted from helpful comments from Richard Bird, Michael Cragg, Aloysius Siow, and workshop participants at the University of Toronto; and from research assistance by Yaw Asamoah-Duodo. The errors are my own.
1. INTRODUCTION

Big-city governments are a fact of American life. Over 7.3 million people live in the 309 square miles in which New York City's government provides municipal services, and in 1989 74% of American employment in the fur goods industry, 35% of the American employment of securities brokers and dealers, and 22% of the American employment of entertainers took place within these boundaries [Vilain, 1991]. Even the 25th largest city in 1992, Austin TX, had 492,000 people in an area of 218 square miles. A political history of twentieth century America without, for instance, John Purroy Mitchell, Fiorello LaGuardia, Frank Hague, Anton Cermak, the Richard Daley's, and Carl Stokes would be seriously incomplete.

Yet economics lacks a theory of what job big-city governments perform and so it has no way of evaluating whether they are doing that job well or poorly. According to the prevailing theory of local public finance—that first enunciated by Tiebout [1956]—big-city governments are either anomalies or mistakes. Bradford and Oates [1975, pg. 55], for instance, compare the presence of an "incurable central city" to the presence of an incurable monopoly.

Tiebout's theory, roughly speaking, combines a normative conclusion that metropolitan areas ought to have a large variety in the types of local public goods being produced with a positive insistence that what ought to be is pretty close to what actually is. Large jurisdictions seem to preclude or at least reduce this variety, and so writers in the Tiebout tradition cannot explain why they exist.
What's wrong here? Jurisdictions and services are not the same things. Tiebout's is a theory about local public goods, not about local governments. Just as General Motors produces both Geo's and Cadillac's and supermarkets sell both yogurt and hamburger, no physical constraint keeps a single government from providing different local public goods to different areas within its jurisdiction. I do not believe that New York City provides the same level of protection against tuberculosis in Little Neck as it does in Central Harlem, and I cannot even think of what it would mean to provide "the same level" of leaf collection services in both Chinatown and Tottenville. A fortiori, I cannot accept a physical constraint that enforces a uniformity I can neither observe nor imagine.

The first problem with Tiebout theory, then, is not simply that it does not explain big-city governments, but that it does not explain local governments at all. Mexico City, the largest city in the world, is run directly by the central government; there is no reason why Scarsdale, Beverly Hills, and Evanston should not be also, if diversity in local public good supply is the only desideratum. Just as a supermarket manager supplies a wide variety of foodstuffs, an upper level government could supply a wide variety of local public goods without the extravagance of local governments. Bewley's [1981] results on endogenous political choice within the Tiebout framework even suggest that the local public goods supermarket might work better than the diversity of governments we observe.

The second problem with Tiebout theory is less fundamental. An explicit premise of this theory is that activities in one location do not impose externalities on other locations. In a sense, this assumption allows us to define relevant jurisdictions in a Tiebout world. Modern
urban economics, however, emphasizes the importance of increasing returns to scale within cities (see, for instance, Henderson [1988] and Glaeser et al. [1992] for contrasting views on the nature of those increasing returns to scale). Increasing returns to scale mean that externalities operating through production are likely to be pervasive within a metropolitan area, and the Tiebout conclusion that a metropolitan area is best organized as a collection of small, varied jurisdictions is open to question.

To be slightly more specific, consider an urban area -- downtown Manhattan in the late 1800's, for instance -- where there are important increasing returns to scale in production. These increasing returns can be realized, however, only if more workers can be brought into the area, but these workers can be brought in only if there is some place for them to live. Roads, sewers, subways, schools and police services, therefore, in the Bronx, Queens, South Brooklyn, and Staten Island give additional Manhattan workers a place to live, and so give Manhattan firms a way to realize increasing returns -- in fact, they make the entire metropolitan economy more productive. This productivity spillover from increasing returns to scale is the pervasive externality to which I have been referring, and it is the reason why the Tiebout conclusion is questionable.

This link with increasing-returns-to-scale production is why big-city governments are not randomly distributed across the countryside, as one would expect if they were anomalies or mistakes. Big-city governments are found where increasing returns to scale are found; they are centered (metaphorically if not geographically) on pieces of geography where economies of agglomeration and urbanization are operating most powerfully. Although the correlation is far from perfect (for reasons I
will explore later), it is not an accident that America's most powerful urban productivity engine, Manhattan, is located in the political jurisdiction with the largest population. New York City is large because Manhattan is productive, not because people like to look at NYPD uniforms or hate making right turns on red (these are just about the only public goods I can think of in New York City that are general throughout all five boroughs). Most people, I think, have an intuitive feel for this relationship between productivity and jurisdictional population, but it does not follow from Tiebout's theory. Part of the task of this paper will be to put this relationship on a rigorous footing, and to understand when to look for exceptions.

Developing a usable theory of big-city governments thus requires two steps: the organizational question of when we would expect to see either unified or separate governments, and the externality question of how increasing returns to scale in production make local public good investments in one location affect land values in other locations. Answering these two question is the goal of this paper; in this way I hope to begin to construct a theory of big-city governments.

The latter question, the one about externalities, is easier and I will attack it first. First I will construct a static model of a metropolitan area of less than optimal size, and show how expanding city services to an unserved region benefits all landowners. Then I will extend the model to a dynamic context and show how the expansion of city services through annexation should proceed.

The cities that are big today got big through annexation -- Jackson [1985, pg. 141] writes that without annexation, "there would now be no
great cities in the United States in the political sense of the term" -- and so any theory about big cities has to include a theory about annexation.

This theory of annexation stands in interesting contrast to Tiebout's theory. Tiebout's theory is about a world of increasing returns in public production but constant returns in private; the emphasis is on consumption. Annexation theory is about a world of constant returns to scale in public production and increasing returns in private production; the emphasis is on production. Tiebout's theory views municipalities as large-scale versions of clubs; annexation theory views them as large-scale versions of plantations or company towns (or universities).

Annexation theory also helps answer one of the small nagging questions of local political economy: why do businesses pay for residential services like recreation, libraries and schools? Businesses seem to be powerful lobbying groups in the other aspects of American political life, especially in municipalities; why do they end up paying a large share of the freight for residential services? The answer from annexation theory is simple: because those residential services are actually part of the production process. The sewers that dispose of the wastes of a firm's workers are as much of an input into its production as the sewers that dispose of its own wastes (for an interesting discussion of this point in nineteenth century England and twentieth century LDC's, see J. Williamson [1990]).

The static and dynamic models of annexation in the next two sections of the paper, however, are largely silent on organizational issues. Implicitly they assume that internalization is the only way to cope with externalities; explicitly they assume there is some fixed cost, "loss of
responsiveness," to being "swallowed up" in a larger entity. Clearly these assumptions need justification -- and qualification.

Section 4, therefore, discusses why internalization is often a better solution for the inter-jurisdictional externalities involved than, say, Coasian contracts or subsidies from higher levels of government. The key issues here turn out to be observability and dynamic consistency.

Section 5 examines the costs of being "swallowed up" in a larger entity. The goal is to formalize the intuition that smaller units of government tend to be more responsive. This section, then, completes part of the Tiebout program by deriving the argument for having large numbers of small governments in a metropolitan area, not just large numbers of different local-public-good packages (or local-public-good-cum-distance packages as in Hamilton [1975]). Only by understanding the advantages of a correctly specified alternative to big-city governments can we understand why and when they are likely to occur. The results of section 5 are also important for a more fundamental reason: they show why we have municipal governments of any kind, big or small, rather than just departments of one large over-arching government.

Section 6 concludes.

2. INCREASING RETURNS TO SCALE AND EXTERNALITIES: THE STATIC VIEW

Consider an open city with all production in an infinitesimally-sized central business district. The only direct input to production is labor, and all output is sold on the world market at price one. Production in the CBD is carried out by a large number of identical, perfectly competitive firms,
but there are external urbanization economies; the larger the population of
the city, the more productive each firm is. Specifically, each firm i's
production function is

\[ f(l_i) = g(L) l_i \]

where \( l_i \) is the amount of labor employed by firm i, L is total city
employment, and \( g(L) \) is an increasing function that reflects economies of
urbanization and agglomeration.

Let \( w \) denote the wage. At a nontrivial equilibrium, obviously,
(1) \[ w = g(L) \]
and each individual firm is indifferent about the size of its labor force,
and makes no profit.

Workers are identical. Each consumes one unit of standard variety
housing, and spends the rest of her income on food, which is available at
price one in the world market. Her utility depends on how much food she
consumes. The world level of utility is \( u \). If \( R(m) \) is the rental cost of
housing \( m \) distance from the CBD, and \( c \) is commuting cost per unit
distance, migrational equilibrium requires
(2) \[ u = w - R(m) - cm, \]
for all \( m \) where workers live. Just as constant returns at the firm level
assures no profit, costless migration at the individual level assures no
worker surplus. Hence any surplus must accrue to landowners, at least in
the long run.

All the land on which workers can live is located along an infinitely
long road that starts at the CBD; the city we are concerned with is long
and narrow. (All the results in this paper would hold for a circular city,
but the algebra would not be so clean.) Each worker's house takes up one
unit of distance; thus precisely m workers can live within m distance of the CBD. There is no alternative use for land along the road. Workers can live only along those sections of the road to which urban services -- paving, street lighting, sewers, city water -- are being supplied. There is no substitute for these urban services.

Let M denote the furthest distance along the road to which urban services are being supplied. Suppose all sites less than M distance have urban services, and none that are further away. Suppose also that all sites with urban services are occupied.

Land market equilibrium requires
\[ R(M) = 0; \]
the most distant worker forces no other worker to live further. From (2), setting \( m = M \),
\[ w = u + cM \]
This equation is essentially a supply-of-labor equation; it says what wage is needed to fill a city size M.

On the other hand, the city's population is also M. Hence from (1):
\[ w = g(M), \]
which is a demand-for-labor equation. It says what wage employers are willing to pay in a city of size M.

Let \( M^* \) denote a city size that equates supply and demand, if such a city size exists, and if the supply curve cuts the demand curve from below, as in figure 1. \( M^* \) is the size the city would have if urban services were supplied costlessly by a benevolent and omniscient planner. (If supply cut demand from above or if demand were always greater than supply, the optimal city would be infinitely big; if demand were always less than supply, the optimal city would be one that did not exist.)
Suppose, though, that for some reason (to be discussed in the next section), for the actual value of M,

\[ g(M) > u + cM \]

The city is "too small;" the demand wage exceeds the supply wage. Demand-for-labor condition (4) will still have to hold -- if firms were making profits from each worker, they would bid against each other for the existing work force and the wage would rise until it equalled marginal product. Workers would not gain from this bidding war, however, since land owners would raise rents. Workers living at the CBD boundary must achieve the world utility level; hence

\[ u = g(M) - R(0) \]

and so

\[ R(0) = g(M) - u \]
\[ R(m) = g(M) - u - cm, \quad m<M \]
\[ R(M) = g(M) - u - cM > 0. \]

Now consider the benefits that accrue from extending urban services by an infinitesimal amount dM. There are two kinds. First is the private benefit to the land owner at that location:

\[ B_p(M) = R(M) \, dM = [g(M) - u - cM] \, dM \]

If \( B_p(M) \) is greater than the cost of urban services, the land owner will install them, but if it is less, he won't. The second benefit is external: increasing M raises productivity g(M), which raises the wage, which raises rent at all locations that are already settled. The social benefit is

\[ B_s(M) = Mg'(M). \]

(This is the same amount that would accrue as profit to firm owners if rents didn't adjust and wages stayed on the supply curve.) Even if
[B_p(M) + B_s(M)] is greater than the cost of urban services, the land owner will not provide urban services if B_p(M) is less than their cost. In this case, though, the other land owners would be better off if they subsidized urban services on land owner M's land.

This is the fundamental externality that drives this paper. Big-city governments are an attempt to deal with this externality. Notice that the externality is present in the same way if we are talking about a group of land owners who constitute a small town, rather than a single infinitesimally small land owner. Nor will the result go away if we relax the assumptions about living conditions by permitting crowding and residence without full urban services: more pleasant conditions will attract more workers at the same wage, increase productivity, and ultimately raise rents everywhere. If all that urban services do is make life more pleasant, that still attracts more workers into the city at the same wage, and so allows more urbanization and agglomeration economies to be realized.

I am not arguing, of course, that the externality arising from increasing returns to scale in production is the only externality that has ever mattered in the determination of city boundaries. Consolidating the harbor district, for instance, was one of the reasons why Staten Island was added to New York City in 1898 [Hammack 1982]. But increasing returns to scale seem to be the dominant externality.

3. A DYNAMIC MODEL OF ANNEXATION

Why would a city be too small? The simplest reason for a city to be too small at a point in time is technological change: it may have been the right size a little while ago, but new kinds of technology have made
expansion possible and desirable. The periods of rapid annexation in U.S. city growth seem to coincide with periods of technological change, and so technological change may be a reasonable force to drive city growth. In this context we can also understand the development of suburban governments.

In a metropolitan area where growth is expected, every small town on the outskirts of a city has a choice. It can either wait for the city to expand, annex it, and pay for the installation of urban services; or it can install the services itself. Annexation carries with itself the one-time cost of \( Q \) per unit distance for having less responsive government. I will show in section 5 why \( Q \) is positive. Roughly speaking, political leaders extract disproportionately greater rents in larger jurisdictions; governments "closer to the people" work better. This idea has a distinguished history in American political thought, and I formalize it in section 5.

An independent town can also decide on its own when to install urban services. Roads, sewers, street lights, water supply systems, fire departments and schools all involve substantial capital costs, and so for this section I will portray the installation of urban services as a one-time capital investment. The actual decision is more complicated than that, and these complications will play a crucial role in section 4.

The type of technological progress that is easiest to model is reduction in transportation cost. In terms of the supply and demand diagrams, a decrease in transportation costs swings the supply curve downward as in figure 2, and increases the size of the city at which supply equals demand. Since the city size is bigger and demand curve is unchanged, lower transportation costs raise the wage also. Since the
relationship between transportation costs and wages is so complex, we will think about the passage of time as simply a force increasing wages.

Consider a landowner at distance $m$ from the CBD trying to choose the optimal time to develop, and trying to decide whether to develop as a suburb or as part of the city. If he develops as a suburb, there will be no subsidy, but no cost from less responsive government either. The timing problem for suburban development is to choose time $T$ to maximize

$$\int_T^\infty R(t,m) e^{-rt} dt - se^{-rT}$$

where $s$ is the capital cost of installing urban services, $r$ is the discount rate, and $R(t,m)$ is the rent at distance $m$ at time $t$. The first order condition for an optimum implies

$$R(T,m) = rs$$

Since

$$R(T,m) = w(T) - u - cm,$$

the first order condition can be written as the requirement that the wage $W_B(m)$ high enough to induce development as a suburb is

$$W_B(m) = rs + u + cm$$

When the wage reaches $W_B(m)$, the land at $m$ will be developed as a suburb, provided it has not been developed as a part of the city already.

For development as part of the city, the reasoning is similar. Assume the subsidy that the rest of the city is willing to pay for development at $m$ is equal to the full value of externality $B_S(m)$, but that the one-time responsiveness cost $Q$ must be incurred. Then by similar reasoning, the wage $W_C(m)$ high enough to induce development as part of the city is

$$W_C(m) = W_B(m) + B_S(m) - rQ$$
Obviously a necessary and sufficient condition for a central city to be surrounded by a suburbs is for $B_S(m)$ to be greater than $rQ$ for small $m$ and less than it for large $m$. For sufficiently large $k$, the Henderson [1988] function
\[ g(L) = k \exp \left( -\frac{\phi}{L} \right) \]
and the simple function
\[ g(L) = k \left( 1 - \frac{1}{L} \right) \]
satisfy this condition. But the same result could also be achieved if, as seems reasonable, $Q$ (the cost of being swallowed up) were a sufficiently rapidly increasing function of city size.

Thus the boundary of the city will end up at $M$ as defined by
\[ B_S(M) = rQ. \]
Inside $M$, land will be developed as part of the city when transportation cost fall low enough that the prevailing wage rises to $W_C(m)$; beyond $M$, land will be developed as suburbs when the prevailing wage rises to $W_B(m)$. The city will keep annexing until its boundary reaches $M$, and during this time all development will take place within city borders. Once the boundary reaches $M$, annexation stops and development is solely a suburban phenomenon. Development never leapfrogs in this simple model; new houses are always built right next to old ones.

Since $B_S(m)$ is a decreasing function of $m$ in the neighborhood of the eventual urban-suburban boundary, the function $W_S(m)$ rises more steeply than the function $W_C(m)$ in this neighborhood. If wages rise at a constant rate over time, the rate of metropolitan growth must slow after development hits the suburban boundary. Because the $W_S(m)$ curve is steeper than the $W_C(m)$ curve, larger wage increments are needed to spur the same amount of development. \textit{Ceteris paribus}, metropolitan areas
where the central city has reached its eventual boundary will be growing more slowly than metropolitan areas in which the central city is still annexing.

4. WHY ANNEX?

Annexation is not the only way to handle externalities. One obvious alternative, for instance, would be for the city landowners to enter into a contract with the suburban landowner to subsidize urban services in the suburb. My neighbors’ car alarms are a source of significant externalities for me, but I am not therefore contemplating making my neighbors part of my family. Why is annexation such a popular way of dealing with the externalities of urban development?

I can think of two chief reasons. The first is future development. Land at distance \( m \) might be at the urban border now, but in a few years the border will move out to \( m' > m \). Landowners at \( m \) will then gain from urban services and subsequent development at \( m' \) -- whether or not they subsidize those services. The external benefits from development at \( m \) are a public good for every landowner closer in. Annexation makes it more difficult for annexed landowners to free-ride on future development.

The second reason is ex post opportunism -- the classic O. Williamson [1983] rationale for organizational integration. If urban services and development were simply a matter of a one-shot public investment as we portrayed it in section 3, ex post opportunism could not arise. The model in that section, however, omitted two features of urban development that matter a lot for organization.

The first is that increasing a metropolitan area’s labor force requires private investment as well as public. Houses have to be built as
well as roads and schools; plumbing as well as sewer and water systems. The second omitted feature is that continuing habitability requires continuing expenditures: water and sewer plants have to operate, police have to patrol, teacher have to teach, firefighters have to be ready to respond, street lights have to be lit, roads have to be cleaned and the snow on them plowed, and all of these systems have to be repaired when something breaks.

Together these two features create a problem of ex-post opportunism. Suppose I build a house in Staten Island, relying on Manhattan's government for sewer and water service. Manhattan's interest is more workers for the CBD. Once I have built the house, it is worthless without water and sewer. So Manhattan's government can raise my water and sewer rates until the sum of the other operating costs and water and sewer rates is just equal to rent; then I will continue to use the house as a home for CBD workers, but I will realize no return on my investment. Thinking about this scenario, I will not build the house unless Manhattan's government can commit not to act opportunistically after I have built it.

Putting such a commitment in writing is difficult. How could the central city government commit itself, for instance, to a level of policing? Notice that a contract to provide a fixed dollar amount to the suburban town would not be acceptable to the central city: the suburb would use the money for its own consumption, not for expanding urban services and raising the metropolitan area's population.

Giving Staten Island a vote in citywide elections, in the city council, and on the Board of Estimate, though, makes it less likely that the citywide government will exploit Staten Island opportunistically. Direct
provision of the urban services assures that they raise metropolitan population. Annexation thus helps mitigate ex post opportunism by the central city government as well as future free-riding by the suburban government.

Neither of these problems would arise, however, if there were neither a central city government nor a suburban one. A stronger upper level government that controlled all of the metropolitan area *ab initio* would avert all of these problems. It could decide the proper times to deliver urban services to each part of the growing metropolitan area. In the U.S., city and town governments are "creatures of the state"—their powers are derived totally by delegation from state governments. The contracting and externality problems I have been describing heretofore would not have happened if this delegation had not occurred. Why then were these creatures created? That is one of the questions I examine in the next section.

5. THE ADVANTAGES OF SMALL GOVERNMENTS

No matter what size a jurisdiction is, the problem of government accountability remains. Indeed, in the paradigmatic economic model of accountability -- the principal-agent problem -- the jurisdiction's population is only one -- a single principal trying to control a single agent. Local government is replete with agency problems: citizens want their snow plowed, their garbage picked up, their streets patrolled and lit, their fires extinguished, but it is very difficult for them to tell whether public officials are putting their best efforts into accomplishing those ends or relaxing on the job, hiring their incompetent relatives, and
amassing publicity and campaign funds for attempts at higher office, and
enriching themselves from kickbacks from contractors.

One would think, therefore, that bigger jurisdictions would be better
because their citizens' experiences, collectively, would contain more
information about officials' efforts. Bigger samples are more
informative. Indeed, this relationship can be demonstrated rigorously
when the government has only two possible actions -- working or shirking
-- and each citizen receives a signal from the same (symmetric unimodal)
distribution. In [1990] I showed that under these circumstances majority
rule was the unique nonmanipulable mechanism, and that the payment
required to make the official work rather than shirk was a decreasing
function of the square root of the number of citizens.

This result holds, however, only when citizens have identical
concerns about government action and all citizens' signals are drawn from
the same distribution--for instance, when the public official is providing
a public good to all of them. An example is a park used by everyone in the
town: the superintendent either works or shirks and everyone notices in
random fashion how well the park is maintained; but contingent on the
superintendent's either working or shirking, the quality of every citizen's
experience in the park can be thought of as a random variable drawn from
the same distribution.

Suppose, though, that the park had two parts used by different
people -- woods and fields, for instance. The superintendent could either
work or shirk in each part. Then the signals that woods-users received
would not necessarily be drawn from the same distribution as the signals
that the field-users received, since the superintendent could work in one
part and shirk in the other.
Similarly and more relevantly, if the Mayor of New York City can make separate decisions about policing Brooklyn and policing Staten Island -- and we have argued in the introduction that he can -- then Brooklynites and Staten Islanders can receive signals about policing that are drawn from different distributions. Bigger samples would not be more informative. On the contrary, incorporating the two boroughs into one polity would hurt their citizens.

To show this, we need to formalize. Consider two sets of voters B and I (for Brooklyn and Staten Island respectively). Assume that both |B| and |I| are large and odd. Each Brooklyn voter receives a signal \( x \) about the quality of city services in Brooklyn, and cares only about the effort the Mayor makes in Brooklyn. Each Staten Island voter receives a signal \( y \) about the quality of municipal services in Staten Island, and cares only about the Mayor's efforts in Staten Island. These are the principals.

The agent is the Mayor, who decides an action to take in each borough. In each borough he can either work or shirk. In the absence of rewards he would prefer to shirk in both boroughs. Specifically, he endures a private cost of \( c_B \) if he works rather than shirks in Brooklyn, \( c_I \) if he works rather than shirks in Staten Island. We assume that, all things considered, \( c_B \) and \( c_I \) are low enough, and the public benefits of working high enough, that the efficient outcome is for the Mayor to work in both boroughs, and ask about mechanisms that implement that outcome. To simplify further, we assume the Mayor has no alternative employment, so we need not include a participation constraint.

Whether the Mayor works or not affects the distribution of signals voters receive. When the Mayor works in Brooklyn, the signals \( x \) that Brooklyn voters receive are i.i.d. random variables drawn from a
symmetric and unimodal distribution with mean and mode at $\mu > 0$. Denote the pdf of this distribution as $f_w(.)$. When the Mayor shirks in Brooklyn, Brooklyn voters' signals are i.i.d. variables with pdf $f_s(.)$, where $f_s(.)$ is simply a leftward translation of $f_w(.)$:

$$f_s(x) = f_w(x + 2\mu).$$

Hence the mean and mode of Brooklyn voters' signals is at $(-\mu)$ if the Mayor shirks in Brooklyn.

Staten Island voters' signals have the same properties, mutatis
mutandis. The pdf for $y$ is $f_w(.)$ if the Mayor works in Staten Island, and $f_s(.)$ if he shirks.

The revelation principle [Gibbard (1973), Myerson(1985)] assures that we need consider only direct mechanisms -- mechanisms where the voters report their experiences to the center, which then orders payments as a function of these reports. Let $(X,Y)$ denote a vector of these reports; the dimension of this vector is $|B| + |I|$. A mechanism $M$ is a function that assigns to each vector $(X,Y)$ a non-negative payment $M(X,Y)$ to the Mayor. For any Staten Island vector of signals $Y$ let $p(Y)$ denote the set of payments that might be made with that signal:

$$p(Y) = \{ p | (\exists X) M(X,Y) = p \}.$$

For any pair of a payment $p$ and a Staten island vector $Y$, let $Q(p,Y)$ denote the set of Brooklyn vectors that give rise to payment $p$ in the presence of Staten Island vector $Y$:

$$Q(p,Y) = \{ X | M(X,Y) = p \}.$$

For any vector of Brooklyn signals $X$, let $F_w(X)$ denote the probability of that vector if the Mayor works in Brooklyn, and $F_s(X)$ the probability if he shirks. Similarly, let $G_w(Y)$ denote the probability that $y$ will be the
vector of Staten Island signals if the Mayor works, and $G_s(Y)$ the probability if he shirks. By independence, then, the probability of observing $(X, Y)$ if the Mayor, say, works in Brooklyn and shirks in Staten Island is $F_w(X)G_s(Y)$.

The Mayor will work in Brooklyn if and only if the expected return from doing so is greater than the expected return from shirking. The expected payment under mechanism $M$ to the Mayor when he works in Brooklyn and Staten Island is:

$$\int G_w(Y) \int_p \int_{X \in Q(p, Y)} p F_w(X) dX dp dY = P_{ww}(M)$$

Define $P_{ws}(M)$ and $P_{sw}(M)$ analogously. Since we are interested in mechanisms that induce the Mayor to work in both places, the Brooklyn part of the mechanism must be designed to induce the Mayor to work in Brooklyn when he is also working in Staten Island. Hence the relevant condition for working in Brooklyn is

$$P_{ww}(M) - c_B - c_l \geq P_{sw}(M) - c_l$$

or

$$(5) \quad P_{ww}(M) - P_{sw}(M) \geq c_B$$

and the relevant condition for Staten Island is

$$(6) \quad P_{ww}(M) - P_{ws}(M) \geq c_l$$

A mechanism $M$ is called wasteful for Brooklyn if (5) is a strict inequality; $M$ is wasteful for Staten Island if (6) is a strict inequality. Suppose that a mechanism is wasteful for Brooklyn. Then any Brooklyn
voter could propose an alternative mechanism that paid the Mayor less in certain states and he would still work. All Brooklyn voters, since they pay less and receive just as good services, would agree to this new mechanism. But a wasteful mechanism need not lead to a formal revision. If the wasteful mechanism were not revised, some or all of the Brooklyn voters could extort a bribe from the Mayor for the privilege of being mayor, and no one else would object (except out of jealousy) or some group of Brooklyn voters could sometimes report their experiences untruthfully if they were not bribed not to. I do not need to be explicit at this point about what happens when a mechanism is wasteful; instead I will look for mechanisms that are not wasteful.

Wastefulness, though, is only a sufficient condition, not a necessary one, for instability, bribery, and chicanery. Call a mechanism M* a *Brooklyn-replica* of mechanism M if, whenever M*(X, Y)=p for any (X, Y) and p, there is some Z such that M(Z, Y)=p. Brooklyn voters, acting on their own, could change M into any of its Brooklyn replicas, simply by reporting Z whenever their experiences were really X. A mechanism M is *Brooklyn-manipulable* if any of its Brooklyn-replicas is wasteful. If a mechanism were Brooklyn-manipulable, someone could convincingly tell the Brooklyn voters: "Report your experiences to me instead of to the center. If you report X to me, I will report Z to the center, and so forth. In this way we will make a Brooklyn-wasteful mechanism M*. Then we can extort a bribe from the Mayor, who will still work, or just simply reduce our payments, and all of us will be better off."

Define Staten-Island-manipulability analogously. A mechanism is *nonmanipulable* if it is neither Brooklyn- nor Staten-Island-manipulable. The primary result of this section is that decentralized majority rule is
essentially the only nonmanipulable mechanism. By decentralized majority rule I mean a mechanism where each voter votes "yes" if and only if her experience was positive; and the Mayor receives one amount $p_B$ if the majority of Brooklyn votes are "yes" and a majority of Staten Island votes are not; another amount $p_I$ if the opposite occurs; and the sum $(p_B + p_I)$ if majorities in both boroughs votes yes. Since the consolidated City of New York is not governed by decentralized majority rule, whatever mechanism it employs must be manipulable.

To derive the primary result about nonmanipulability, rewrite the left-hand side of (5)

\[
P_{ww}(M) - P_{sw}(M) = \int G_w(Y) \int p \int \delta(X) \, dX \, dp \, dY \quad \text{for } X \in Q(p, Y)
\]

where

\[
\delta(X) = F_w(X) - F_s(X).
\]

If $M$ is nonmanipulable, it cannot be Brooklyn-wasteful,

\[
P_{ww}(M) - P_{sw}(M) = c_B
\]

and so for any Brooklyn-replica $M^*$

\[
P_{ww}(M^*) - P_{sw}(M^*) \leq c_B
\]

Thus for any Brooklyn-replica $M^*$

(7) \quad P_{ww}(M) - P_{sw}(M) - [P_{ww}(M^*) - P_{sw}(M^*)] \geq 0.

Note that if $M^*$ is a Brooklyn-replica of $M$, $p(y) = p^*(y)$, and so (7) becomes (8)
\[ \int G_w(Y) \{ \int \delta(X) \, dX \, dp - \int p \\delta(X) \, dX \, dp \} \, dY \geq 0. \]

Suppose some mechanism has the following property:

**property (*)**: for every \( Y \) and its associated \( p(Y) \), \( Q(p,Y) \) maximizes

\[
\int p \, \int \delta(X) \, dX \, dp. 
\]

Then that mechanism is surely not Brooklyn-manipulable, because the expression in curly brackets in (8) is always non-negative, for any Brooklyn replica.

I claim that decentralized majority rule has property (*). First suppose \( p(Y) \) has precisely one positive element. Then the problem of choosing \( Q(p,y) \) to maximize (9) is simply the problem of finding a set of Brooklyn experience vectors that maximize

\[
\int \delta(X) \, dX. 
\]

This is the simplest, one-dimensional election problem, and it was solved in O'Flaherty [1990, proposition 1]. The solution is simple: \( Q(p,y) \) should include all these vectors and only those vectors where a majority of Brooklyn voters have positive experiences. (The intuition is that this mechanism makes the median voter's experience decisive, and the median is the most powerful order statistic. The mean is more powerful, but is not incentive-compatible.) Denote this solution as \( Q^m \).

\[ Q^m = \{ X | \{|x_i \geq 0|\} > |B|/2 \}. \]
If \( p(Y) = \{0\} \), then \( Q^m \) still maximizes (9) trivially; any set would do so. By construction, \( p(Y) \) cannot be empty.

Finally, suppose \( p(Y) \) contains more than one positive element. Let \( p^* > 0 \) denote the largest element of \( p(Y) \) and \( p' \) denote the smallest; it may be zero. Then (9) is maximized by setting \( Q(p^*, y) = Q^m \) and setting \( Q(p', y) \) equal to the complement of \( Q^m \).

To show that this solution is unique, simply note that any other mechanism, when compared with a mechanism with property\((\ast)\), would have the expression in curly brackets negative for some \( Y \); and could not have the expression in curly brackets positive for any \( Y \); and so would be Brooklyn-manipulable.

Thus if a mechanism is not Brooklyn-manipulable, it must work like this: for any \( Y \), pay the highest amount possible if the Brooklyn experiences are in \( Q^m \), and the lowest amount possible if they are not. By similar reasoning, mechanisms that are not Staten-Island-manipulable must act the same way: for any \( X \), pay the highest amount possible if the majority of Staten Island voters' experiences are positive, and the lowest amount possible if they are not. Hence if a mechanism is not manipulable, it can have only four levels of payment to the Mayor: nothing, if majorities in both boroughs have negative experience; an amount \( p_I \) if a majority of Staten Islanders have positive experiences and a majority of Brooklynites don't; an amount \( p_B \) if the opposite occurs; and \( (p_B + p_I) \) if majorities in both boroughs have positive experiences. This is precisely how I have defined decentralized majority rule.

Decentralized majority rule could be implemented if Brooklyn and Staten Island had separate mayors and separate elections (it wouldn't matter if the same person held both jobs; many small towns in New
Jersey, for instance, share tax assessors and health officers). It could not be implemented if Brooklyn and Staten Island were part of one city that held city-wide elections. Thus the price of consolidation is electoral manipulability.

How high is this price? Without an explicit theory of how manipulable mechanisms get manipulated I cannot say. If communication, information, and coordination were costless, manipulations could be set up instantaneously and mechanisms would automatically become nonmanipulable. The formal rules of an election would be irrelevant. But manipulation is expensive, and proficient manipulators no doubt realize rent from their activities. There may be efficiency losses as well because some outcomes that are expensive for the Mayor cannot be implemented.

Reducing the scope for manipulation, therefore, is why small jurisdictions make sense. The costs of governmental consolidation that we appealed to in section 3 are real.

6. CONCLUSION

Tiebout's theory provides many important insights into how local public goods are and should be provided in metropolitan areas. But it says little about either big cities or the organization of government. In this paper I have tried to appeal to increasing returns to scale in production and to information in order to explain big city governments.

The theory I have sketched is roughly in accord with the stylized facts of big city development. The largest political jurisdictions tend to be those that include areas where large economies of agglomeration and urbanization are being realized. Those jurisdictions became large by
aggressively annexing surrounding towns and areas--before those areas were thickly settled (Staten Island had a population of 391,000 in 1992, but the vote in favor of consolidation in 1894 was 5531 to 1505; Queens with a 1992 population of 1,951,000 voted 7712 for and 4741 against [Hammack, 1982, p.206]). Developed areas do not get annexed--witness Minneapolis and St. Paul or the long coexistence of Manhattan and Brooklyn. Rapid annexation takes place during the times when metropolitan areas are growing most quickly, and then stops. Metropolitan areas where cities are not annexing tend to grow less quickly. The greatest support for annexation comes from downtown business interests and the owners of outlying land (see Teaford [1979] and Hammack [1982]). All of these stylized facts are in accordance with the theory I have set out in this paper.

Many other puzzles remain. I do not know why neighborhoods that get annexed tend to stay annexed--why secession is so rare in the annals of local government. Perhaps subsidies continue; perhaps capital structure is built in such a way that neighborhoods are locked in; perhaps powerful political forces are realizing rent from the manipulations that centralized government permits. This question is particularly important for understanding cities like Newark where the original rationale for annexation has all but disappeared (in 1990, only 46.6% of the workers who lived in Newark went to work in Newark [U.S. Bureau of the Census, 1993]).

I do not know why state boundaries hardly ever change. Since 1800, only two major state boundary changes have occurred (both in connection with the Civil War -- the additions of Maine and West Virginia). State boundaries prevented the New York and Philadelphia suburbs in New Jersey
from being annexed; that's why Hudson, Bergen and Camden counties in New Jersey have such a plethora of geographically tiny communities. But why those state boundaries did not change I do not know.

Free-riding by businesses--or more precisely, by owners of land on which businesses are located--is another area that this paper has failed to address, but which a full theory of local governments need to consider. In this paper I have forced all businesses to locate in the CBD (where the increasing returns to scale are realized), and since the CBD had an area of measure zero, CBD land contributed nothing to the subsidy for development. Considering a CBD with a positive measure of land and land as an input to production would be a mindless extension of the model, and would present not real difficulties.

Suppose, though, that there were a few other spots in the metropolitan area outside the CBD where a small number of businesses could also locate. Assume further that most of the increasing returns are metropolitan-wide. Then businesses in these outlying islands could benefit from residential expansion, but could avoid paying for it by incorporating as separate jurisdictions--a classic free rider situation. Teterboro, New Jersey is the most frequently cited example of a business free-riding jurisdiction, but Asamoah-Duodo [1994] lists East Newark (a town of six blocks around the Clark thread mill), Harrison, Hillside, Belleville, and Irvington as industrial towns that surrounded Newark in the late nineteenth century. Gordon [1977] also emphasizes business free-riding in his story of why annexation ended in the east.

A final weakness in this paper is its treatment of government output as single good. Governments provide many different goods--fire protection, police, sewers, water, schools, libraries, parades, holiday
celebrations—and no physical constraints requires that the same
government provide all of them to a location. Annexation need not be a
80-81] have argued that the rise of special purpose multi-jurisdictional
districts, especially those for sewers, at the turn of the century was in
part responsible for the cessation of annexation in the East and Midwest
at about that time. But in fact there are good organizational reasons to
consolidate public services that affect the same group of people: the same
non-manipulation criterion that argues for separate elections when
separate groups of people are affected by distinct decisions also argues
for a single election when the same group of people are affected by
distinct decisions. I state and demonstrate this proposition rigorously in
[1995]. Towns should be small, but they should have many functions.

Thus in most cases special districts act like subcontractors to
municipalities. A town may buy its firetrucks from La France, but voters
hold local officials responsible for the performance of those firetrucks,
not La France. Thus we do not think of the United States as a consolidated
"firetruck district" governed by La France. Municipal accounting systems
treat payments to special districts like Passaic Valley Sewerage
Commissioners the same as payments to La France; this is probably the
right way to think about the Passaic Valley Sewerage Commissioners.
Like La France, special districts are generally a way of helping local
elected officials do their jobs; they may do so well or poorly. (Multi-
jurisdictional public utilities like water companies and cable television
systems should probably be treated the same way; so should the Compton
Plan.)
Why should anyone care how big cities are governed? Public health, physical security, and education are areas of great concern. Big-city governments provide these services to large numbers of people -- and especially, to large proportions of poor people and African-Americans. Only with a theory about why big-city governments exist and how they operate can we begin to ask how well they do this job and whether alternative arrangements could do it better.
Note

1. Only land owners gain from development because in this model, as in almost all the literature on monocentric open cities, land owners are the only people who can gain (or lose) from anything. Slightly more generally, any owner of a location-specific asset in fixed supply can gain, and so has an incentive to promote development. Thus owners of utility franchises were often among the more aggressive proponents of annexation. (See, for instance, Asamoah-Duodo [1994] for the case of Irvington, N.J.)
References


Bruce Hamilton [1975]: "Zoning and Property Taxation in a System of Local Governments," *Urban Studies* 12, 205-211.


Jeffrey Williamson [1990]: *Coping with City Growth during the British Industrial Revolution*, New York, Cambridge University Press.

Figure 1:

Supply: \( w = u + cM \)

Demand: \( w = g(m) \)

M

Figure 2:
The following papers are published in the 1994-95 Columbia University Discussion Paper series which runs from early November to October 31 (Academic Year). Domestic orders for discussion papers are available for purchase at $8.00 (US) each and $140.00 (US) for the series. Foreign orders cost $10.00 (US) for individual paper and $185.00 for the series. To order discussion papers, please send your check or money order payable to Department of Economics, Columbia University to the above address. Be sure to include the series number for the paper when you place an order.

708. Trade and Wages: Choosing among Alternative Explanations
   Jagdish Bhagwati

709. Dynamics of Canadian Welfare Participation
   Garrey F. Barret, Michael I. Cragg

710. Much Ado About Nothing? Capital Market Reaction to Changes in
      Antitrust Precedent concerning Exclusive Territories.
      Sherry A. Glied, Randall S. Kroszner

711. The Cost of Diabetes
      Matthew Kahn

712. Evidence on Unobserved Polluter Abatement Effort
      Matthew E. Kahn

713. The Premium for Skills: Evidence from Mexico
      Michael Cragg

714. Measuring the Incentive to be Homeless
      Michael Cragg Mario Epelaum

715. The WTO: What Next?
      Jagdish Bhagwati

716. Do Converters Facilitate the Transition to a New Compatible Technology?
      A Dynamic Analysis of Converters
      Jay Phil Choi
1994-95 Discussion Paper Series

717. Wealth Effects, Distribution and The Theory of Organization
   Patrick Legros, Cornell University
   -Andrew F. Newman, Columbia University

718. Trade and the Environment: Does Environmental Diversity Detract from the Case for Free Trade?
   -Jagdish Bhagwati and T.N. Srinivasan (Yale Univ)

719. US Trade Policy: Successes and Failures
   -Jagdish Bhagwati

720. Distribution of the Disinflation of Prices in 1990-91 Compared with Previous Business Cycles
   -Philip Cagan

721. Consequences of Discretion in the Formation of Commodities Policy
   -John McLaren

722. The Provision of (Two-Way) Converters in the Transition Process to a New Incompatible Technology
   -Jay Pil Choi

723. Globalization, Sovereignty and Democracy
   -Jagdish Bhagwati

724. Preemptive R&D, Rent Dissipation and the "Leverage Theory"
   -Jay Pil Choi

725. The WTO’s Agenda: Environment and Labour Standards, Competition Policy and the Question of Regionalism
   -Jagdish Bhagwati

726. US Trade Policy: The Infatuation with FTAs
   -Jagdish Bhagwati

727. Democracy and Development: New Thinking on an Old Question
   -Jagish Bhagwati