Universidad de los Andes

Facultad de Economía

Should drug policy be aimed against cartel leaders? Breaking down a peaceful equilibrium

Advisor: Daniel Mejía

Presented by: Juan Camilo Castillo (200621582)

July 31, 2013

Abstract

Experience from the last decade in Colombia and Mexico suggests that violence increases when governments achieve their objective of beheading and fragmenting drug trafficking organizations (DTOs). In this paper I provide a theoretical framework to understand this behavior. Drawing elements from industrial organization, I model DTOs as firms that collude by not attacking each other in order to increase their profits. DTOs always collude when they interact repeatedly; thus, previous analyses focusing on a static Nash equilibrium miss an important part of the dynamics between DTOs. I show that a peaceful equilibrium arises if there are only a few DTOs that care enough about the future. Policies resulting either in a larger number of DTOs or in more impatient leaders increase violence between DTOs without reducing supply. On the other hand, policies that reduce the productivity of DTOs, without directly attacking cartel leaders and fragmenting DTOs, are more desirable since they can curb supply, although this comes at the cost of increased violence if the elasticity of demand is below a certain threshold. I calculate this threshold, which is a refinement of the value suggested by Becker et al. (2006) for consumer markets.

Keywords: War on Drugs, Illegal Drug Markets, Violence, Supply Reduction
JEL Classification Numbers: D74, K42

1 Introduction

The illegal drug trade has shown an exceptional capability to transform itself according to the conditions it must face. Governments have tried to attack drug trafficking organizations (DTOs) with various methods, resulting in traffickers finding new ways to operate, both locally and globally. Whenever authorities are able to eliminate one major trafficking route, a new one arises to replace it. And whenever one form of organization is effectively suppressed, new types of cartels arise.

Among the various mutations DTOs have gone through, a perplexing one has happened when local governments turn to strategies that end up decentralizing control of the drug trade. The idea, based on the premises of the war on drugs led by the U.S., is that the illegal drug trade cannot be eliminated for good if large, powerful DTOs persist. The outcome has been, however, quite different from what governments wanted: drug trade continues, led by former bosses' lieutenants, with an important increase in violence as an unintended consequence. The surge in violence is driven by an increase in the number of drug traffickers killed by other drug traffickers. The death of cartel leaders gives rise to a chaotic state with a void of power, in which DTOs increase the intensity of fights between each other, and in which members within each cartel fight to become the new leaders.

Some clear examples of this phenomenon come to mind. In the first half of the 2000s, a few powerful Mexican cartels increased their dominance of the illegal drug trade, smuggling cocaine, marijuana, methamphetamine and heroin into the U.S. Surprisingly, the levels of violence were low in comparison with other Latin American nations, despite DTOs having gained an amount of power that Mexicans were not willing to accept. This resulted in president Felipe Calderón being elected in 2006 after a campaign based on the promise of frontal war against DTOs. Keeping his word, he started his term with large-scale military actions that continued until the end of his sexennium. Whereas the amount of cocaine crossing the U.S. border did not change significantly, violence started to increase year after year, to the point that the homicide rate in 2010 was more that twice the homicide rate in 2006 (Guerrero, 2011; Castillo et al., 2013).

Another example could be seen in some regions in Colombia. During the early 1990s some paramilitary groups had formed in order to defend populations from left-wing guerrillas like the FARC[1] and the ELN[2]. These groups started forging alliances with drug traffickers, and after a few years paramilitaries and DTOs were undistinguishable from each other. Their peak of power came after they united under the single leadership of Carlos Castaño in the late 1990s under the name

[1] Fuerzas Armadas Revolucionarias de Colombia

[2] Ejército de Liberación Nacional

of the AUC[3]. Regions under their full control experienced a degree of peacefulness seen in few regions in the country at the time. The AUC ended around 2005, when they agreed to demobilize in a treaty with president Álvaro Uribe's government. Just as in Mexico, the subsequent void of power did not lead to a decrease in the amount of drugs produced and trafficked. Instead, multiple small bands emerged to fill up the position of control formerly held by the AUC. These groups, called bacrims (short for criminal bands in Spanish) have led multiple fights over the control of routes, breaking the previous state of calm (Camacho, 2009, 2011).

Although multiple observers have described this kind of behavior (again, see Guerrero, 2011; Castillo et al., 2013; Camacho, 2009, 2011), it has not been described satisfactorily from the economic theory of conflicts, which traditionally focuses on the static interaction between DTOs. The idea that fragmentation of cartels increases violence is not new (O'Neil, 2009), but few analysts talk about the possibility of a peaceful equilibrium like the one observed before governmental interventions in Colombia and Mexico. The purpose of this paper is to provide a theoretical framework that explains the behavior of DTOs when they interact repeatedly. Once they start caring about the future, DTOs collude, following a tacit treaty requiring them not to attack each other, while each one controls a given fraction of the drug trade. They are then able to receive larger profits, since they do not have to spend resources in the conflict, and they have no losses from dead personnel and destruction. However, just as in the theory of collusion in industrial organization, every individual DTO has incentives to deviate. In this context deviating means betraying others by attacking them in order to seize a larger portion of the drug trade, taking advantage that other DTOs are not expecting a betrayal. Therefore, a peaceful equilibrium can only be sustained if DTOs care enough about the future that they prefer the collusive equilibrium to hold for a long period. Additionally, the number of DTOs must be low. If this is not the case, DTOs attack each other: war erupts, and DTOs fight for the control of the drug trade.

The model allows me to compare two approaches to drug policy that have been followed by governments. The first one, which I call enforcement, aims against the operations of DTOs, with the objective of reducing their efficiency as drug traffickers. The second approach is to attack drug cartel leaders frontally, as Calderón did in Mexico. Many analyses had compared the efficiency of government actions at different stages of the production chain of drugs until they reach final consumers, but no study had compared enforcement activities with direct attacks against cartel bosses. The usual criterion to evaluate policies has been effectiveness in reducing supply, despite calls for a more complete treatment: the assessment of policies should be based on how they increase or decrease the total harm caused by drugs, which includes violence related

[3] Autodefensas Unidas de Colombia

to illegal markets (Caulkins and Reuter, 1995). As an answer to these calls, I follow a more global approach, since I evaluate policies in terms of their effect both on consumer nations, who want to curb supply, and on trafficking nations, who want to minimize violence.

Becker et al. (2006) point out that under inelastic demand reducing supply increases the market size, thus increasing the harmful effects of drug markets in consumer nations. This has led to the widespread but imperfect notion that, if demand is inelastic, reducing supply induces higher levels of violence between DTOs. I improve this analysis by considering DTOs' costs as well, and conclude that if demand is more inelastic than certain threshold, enforcement increases violence. The new threshold means that it is easier for enforcement to increase violence than previously thought. Then I show that enforcement, defined as policies that undermine DTOs' productivity, is somewhat beneficial since it reduces supply, but this usually comes at the cost of increased violence because of the new threshold. On the other hand, attacks on drug leaders increase violence without any effect on supply. Thus, only enforcement is effective if the objective is supply reduction, but authorities should be aware of the potential damage it brings to trafficking nations.

Finally, I show that my model fits the Colombian and Mexican cases. As the Colombian government succeeded in signing a demobilization treaty with the AUC, they drastically increased the number of DTOs operating, breaking down the peaceful equilibrium and increasing violence. I also show that the Mexican government's strategy of beheading DTOs led to cartel bosses being more short-sighted, and to an increase in the number of independent DTOs, inducing brutal wars between them.

2 Literature Review

This paper models the illegal drug trade as a series of conflicts[4] that take place from the earlier stages of drug production until consumption. A conflict is a situation in which a number of actors engage in a zero-sum game with the aim of obtaining some prize by investing some effort or resources[5]. The first conflict occurs when rival DTOs engage over the control of routes used to transport routes to their destination, and a second conflict occurs when DTOs try to transport drugs through routes under their control while government forces try to seize these drugs before they reach their destination.

An important branch of the economic theory of conflict focuses on illegal drug markets, their effect on society, and actions that governments can take against them. Becker et al. (2006) make

[4] Garfinkel and Skaperdas (2007) provide a good survey of literature on the economic theory of conflicts.

[5] This simple definition of conflict is different from the much more specific idea of an armed conflict, broadly used in political science, that involves a number of groups fighting for political control of some territory or nation.

one of the main contributions. They argue that the elasticity of demand determines the effectiveness of enforcement activities: if demand for drugs is inelastic, as empirical evidence suggests it is, enforcement turns out to be very ineffective since very large expenditures only cause a small reduction in the deleterious effect drugs have on society. This happens because decreasing the amount of drugs sold actually increases the market size, measured as the price times the quantity. Their work is mainly focused on the retail market in consumer nations under free entry and exit of DTOs, but their conclusions have often been extended to global drug markets, leading to the widespread notion that reducing supply comes with the side effect of increased violence in earlier stages of the production chain. However, their analysis cannot be taken too literally to trafficking markets and violence. I will therefore present a modification of their result that applies to upstream markets.

Until recently, the bulk of the theoretical work on illegal drug markets focused on consumer nations (Reuter and Kleiman, 1986; Lee, 1993; Poret, 2003). In these nations the main objective of public policy is to reduce supply in order to minimize the harmful consequences of drugs on consumers, whereas violence is a relatively mild concern. These works, however, ignore the whole production chain, which starts with crops that are used to produce drugs, then goes through drug-trafficking markets, and finally ends in main consumer markets in the U.S. and Europe. A complete assessment of global drug markets that takes into account all harm caused to society (Caulkins and Reuter, 1995) should include previous stages of the production chain, especially due to the high levels of violence caused in drug trafficking and producing nations.

The last few years have seen growing concern for the situation in the earlier stages of the drug trade. Various authors have focused on cocaine-producing Andean nations, particularly Colombia, and drug-trafficking countries like Mexico and other Central American nations. Grossman and Mejía (2008) build a model in order to compare the relative efficiency of governmental intervention with two different strategies against the cocaine production chain: first, control of land where coca plant is being cultivated, and second, eradication of coca plants and interdiction of the produce of coca crops. A more complete model is analyzed in Mejía and Restrepo (2008), which includes conflict over the control of arable land and conflict over the control of routes. The idea is to evaluate Plan Colombia, led by the U.S. government in order to subsidize the war against drug producers and traffickers in Colombia. They conclude that resources would have been better spent if more efforts from Plan Colombia had targeted the conflict against drug traffickers, and not against drug producers. Some other authors (Chumacero, 2008; Bogliacino and Naranjo, 2012) build general equilibrium models that include various stages in the production chain of cocaine. Mejía and Restrepo (2011) analyze the combination of efforts to fight illegal drugs in producer and consumer countries, with the finding that they are complementary.

Although the previous works focus on upstream markets, they focus on supply reduction, while saying little about the high levels of violence caused in producer and trafficking nations. I therefore take a broader view, by also looking at how different policies increase or reduce bloodshed in trafficking nations. Many works on the chain of production of drugs also follow Becker et al. (2006) and assume that demand for drugs is inelastic, which implies that the size of the drug market increases if governments succeed in decreasing supply. From this, they conclude that competing DTOs then fight for higher stakes, resulting in higher levels of violence.[6] This looks like a possible explanation for the Colombian and Mexican cases I mentioned, in which successful policies result in more violence. But this contradicts the fact that no substantial decrease in supply was seen in either case. Thus, there is no adequate theory to explain them. Existing models have also failed to consider multiple-period interaction between opposing sides beyond being leaders and followers *à la* Stackelberg. I attempt to fill this void by modeling DTOs as agents interacting repeatedly, which opens the possibility of governments succeeding in killing leaders and fragmenting DTOs, while achieving no noticeable decrease in supply and inducing an increase in violence.

I also rely heavily on industrial organization, going in line with a recent trend that attempts to explain DTOs as operating in a complex environment that shares many characteristics with traditional industries. Some works have sought to understand the market structure of DTOs without making explicit references to their violent behavior that leads to conflict: Poret (2003) models the vertical structure, where both the wholesale and the retail stages hold some market power. Poret and Téjédo (2006) and Burrus (1999) model the horizontal structure of drug markets. Some other works model the trafficking industry's particularities: Bardey et al. (2013) propose a model of entry and exit in which governments invest larger efforts on catching more experienced DTOs, and Baccara and Bar-Isaac (2008) study the most efficient organizational structures of criminal groups.

I model drug traffickers in a region as an industry with barriers to entry, in which firms (DTOs) may choose, if it is in their own interest, to collude instead of engaging in free competition, i.e., war. The notion of a market with collusion is based on standard references on industrial organization such as Tirole (1988) and Motta (2004), and on game theory such as Mailath and Samuelson (2006). The work by Abreu (1983, 1986) is especially relevant; he explores how to find optimal punishment strategies in an oligopoly. The novelty of my work is the application of such widespread models from industrial organization to the specific case of violent DTOs that do not

[6] As I already pointed out, this is actually an out-of-context interpretation of Becker's result.

compete on price or quantity, but on the amount of resources they spend in the conflict over the control of smuggling routes. There is another important difference between DTOs and firms in a traditional oligopoly: whereas enforcing treaties is difficult in oligopolies, violent DTOs have the means to enforce tacit treaties, by attacking those who do not follow the terms of the agreement.

3 The model

3.1 Description of the trafficking industry

DTOs are pure drug traffickers, whose sole purpose is to maximize their profit, without any craving for reputation or political control[7]. A fixed number n of DTOs participate in drug trafficking. The number is constant because of barriers to entry: if any new actor tried to enter the market, all incumbent DTOs would join their forces against the newcomer to preclude its entrance. I denote the set of indices for DTOs by I.

The government in the territory is a non-strategic player whose actions are determined exogenously. In real life, governments are clearly strategic actors. However, they pursue multiple goals, such as minimizing supply and minimizing violence, while being subject to a budget constraint. Incorporating the government's preferences in order to determine its strategic behavior involves very strong assumptions about how it weighs its goals. Hence, I assume that the government sets its policy, after which DTOs interact strategically. Given some policy, I will characterize the equilibrium reached, and I will analyze the comparative statics with respect to changes in policy. This will allow me to see how these changes affect the equilibrium of the interaction between DTOs, and how they may help to fulfill the government's goals.

The behavior of DTO $i \in I$ can be divided in two parts. In its *productive behavior*, it buys x_i drugs in producer markets and takes them through an intermediate territory in order to transport them to a consumer market. In order to do so, it must evade the government in the intermediate territory, which spends an amount of resources e in enforcement, in order to seize drugs. DTO i therefore uses two factors of production, an amount of routes R_i and drugs x_i, in order to obtain its final good, drugs in consumer markets, which I will denote by q_i. Hence, DTOs' productive behavior determines the supply of drugs, $Q = \sum_{i \in I} q_i$, which is the total amount of drugs sold by DTOs in the consumer region. It is thus in the best interest of the consumer region to reduce Q. DTOs' operations also involve their *military behavior*, in which they engage others

[7] This is certainly a strong assumption, as many DTOs have clearly shown (Colombian cartels or guerrillas, for instance). However, my purpose is to model their trafficking behavior, and modeling their political behavior would make the task much more complicated, obscuring my main contributions.

in a fight over routes in the intermediate territory. DTO i commits an amount of resources g_i to the conflict. Larger g_i means that it controls more routes. I will use aggregate spending in the conflict, $G = \sum_{i \in I} g_i$, as a proxy for the level of violence. The trafficking region thus wants G to be as low as possible. An additional quantity that will be of interest throughout this paper is $X = \sum_{i \in I} x_i$, the aggregate amount of drugs bought in the producer market.

Modeling both DTOs' productive and military behavior may seem as an excessive complication, since the strategy space becomes two dimensional. But most of the harm on consumer nations is caused by productive behavior, whereas most of the harm on trafficking nations is caused by military behavior. I must therefore analyze both types of behavior if I want to make a global assessment. I will now explain both sides of their behavior in detail.

3.1.1 Productive behavior

DTOs buy drugs in producer markets at a price p_p, and sell them in consumer markets at a price p_c. Each DTO's share of the total market is small, so they have no market power[8]. Thus, they maximize given a fixed level of prices. The whole region, however, may involve an important share of the total drug trade, with the implication that the total amount of drugs going through the trafficking region has an effect on drug prices. The elasticity of demand of drugs in the producer market is ϵ_c. This elasticity corresponds to the price elasticity of the amount of drugs bought from the trafficking region being analyzed[9]. As a simplifying assumption, I assume that prices in the producer market are fixed, which corresponds to an elasticity $\epsilon_p = \infty$. This greatly simplifies the final expressions that I obtain, without losing any important insight. For completeness, I analyze the trafficking industry with $\epsilon_p \in (0, \infty)$ in appendix C.

The government chooses the amount of resources it spends evenly throughout the region in enforcement, e. With enforcement I mean any type of activity aimed against the productive behavior of DTOs. Some examples are seizing drugs in transit, patrolling routes, or seizing submersibles or airplanes. On the other hand, I do not consider capturing or killing leaders or gunmen to be enforcement: such activities disrupt DTOs' military behavior, but not their behavior as drug traffickers.

DTOs choose their course of action based on the level of enforcement. The amount of drugs

[8] As an example, the Herfindahl index for Mexican cartels is around 0.15, suggesting that this is not a very strong assumption. I made the calculation by myself, based on the data from Castillo et al. (2013). As an important caveat: the calculation of each DTOs' share is based on very unreliable data due to the illegal nature of the industry.

[9] If the elasticity of demand to the total amount of drugs is ϵ_c^T, the elasticity of demand to the amount of drugs through the region of analysis is $\epsilon_c = \frac{\epsilon_c^T}{s}$, where s is the fraction of the drugs demanded in the consumer market supplied through the trafficking region.

bought in the producer market, the amount of routes, and the level of enforcement are put together in a production technology that results in an amount $q(x_i, R_i, e)$ of drugs reaching the consumer market. Function q is the same for all DTOs, which means that they are equally efficient in using routes to take drugs from producer to consumer markets. The production function is twice-differentiable, and it is increasing in both factors of production. It is decreasing in the amount spent by the government in enforcement activities ($\frac{\partial q}{\partial x_i} > 0$, $\frac{\partial q}{\partial R_i} > 0$, and $\frac{\partial q}{\partial e} < 0$). Additionally, it is concave in (x_i, R_i)[10], so the marginal productivity of both factors of production is decreasing ($\frac{\partial^2 q}{\partial x_i^2} < 0$ and $\frac{\partial^2 q}{\partial R_i^2} < 0$).

Any increase in enforcement by the government decreases the marginal productivity of both factors of production: if routes are better watched, a lower fraction of the drugs bought at the producer market reaches the final market, displacing the whole production function down, which results in a decrease in both marginal productivities. Formally, $\frac{\partial^2 q}{\partial e \partial x_i} = \frac{\partial}{\partial e}\left(\frac{\partial q}{\partial x_i}\right) < 0$ and $\frac{\partial^2 q}{\partial e \partial R_i} = \frac{\partial}{\partial e}\left(\frac{\partial q}{\partial R_i}\right) < 0$. Thus, enforcement affects the productive behavior through two channels: it decreases the productivity of x_i (which I define as $\frac{q_i}{x_i}$), since $\frac{\partial}{\partial e}\left(\frac{q_i}{x_i}\right) = \frac{1}{x_i}\frac{\partial q}{\partial e}$, and it decreases marginal productivity of x_i, since $\frac{\partial}{\partial e}\left(\frac{\partial q_i}{\partial x_i}\right) = \frac{\partial^2 q}{\partial e \partial x_i}$. The relative importance of these two effects will have consequences in the outcome of enforcement.

Function q has constant returns to scale. This has very important consequences, so I will describe the production technology in an alternative way that shows that this is a sound assumption. DTO i and government forces engage in a conflict over the control of the x_i drugs that the DTO attempts to take to the consumer market. The fraction of drugs that the DTO succeeds in taking to the consumer market (the *survival rate*) is $w_i \in (0, 1)$. The government uses enforcement e to seize a fraction of x_i (reducing w_i), whereas the DTO uses routes R_i to preclude the government from taking its drugs. However, these routes are only effective so long as they are not too saturated. The survival rate is thus $w(x_i, R_i, e)$, where $\frac{\partial w}{\partial x_i} < 0$, $\frac{\partial w}{\partial R_i} > 0$ and $\frac{\partial w}{\partial e} < 0$. The total quantity of drugs that reaches the consumer market is $q_i = w_i x_i$.

As the DTO increases the amount of drugs it attempts to take to the consumer market, it can maintain the survival rate if both x_i and R_i increase in the same proportion: the routes will be equally saturated, and it will be neither easier nor harder for the government to seize drugs. Hence, the survival rate depends on $r_i = \frac{R_i}{x_i}$, the inverse saturation of routes, and $w(x_i, R_i, e) = w(r_i, e)$. Function w is therefore a contest success function (CSF) where the commitment to the

[10] This seems to be an assumption with the sole purpose of enabling the existence of an equilibrium, but it is actually a consequence of the other conditions imposed on the function. The conditions on the first derivatives, on both second derivatives, and on the cross derivatives (which we will soon state) imply that the function is quasiconcave. I also assume that the function has constant returns to scale. Quasiconcavity and constant returns to scale imply concavity.

conflict by the DTO is measured by r_i, and the commitment by the government is measured by e. Since it depends only on $\frac{R_i}{x_i}$, w is homogeneous of degree zero in (x_i, R_i), and q is homogeneous of degree one (it has constant returns to scale). If $r_i = 0$, the DTO has no routes to transport drugs and $w_i = 0$. If, instead, r_i tends to infinity, the DTO has a surplus of routes, so all the drugs it wishes to take to the consumer region will reach their destination and $w_i = 1$. The marginal productivity of r_i is decreasing, i.e., $\frac{\partial^2 w}{\partial r_i^2} < 0$. This implies that $\frac{\partial^2 q}{\partial x_i \partial R_i} > 0$: routes and drugs are complementary production factors[11].

3.1.2 Conflict over routes

There is a continuum of routes with mass normalized to one, i.e., $\sum_{i \in I} R_i = 1$. DTOs engage in a conflict over the control of these routes. DTO i invests an amount g_i, which includes the salaries of gunmen, the cost of guns, losses associated with dead gunmen, etc. At the end of the conflict, the amount of routes held by the DTO is a twice-differentiable function $R_i(g_i, g_{-i})$ that depends positively on its expenditure and negatively on the total amount $g_{-i} = \sum_{j \neq i} g_j$ spent by all other DTOs: $\frac{\partial R_i}{\partial g_i} > 0$ and $\frac{\partial R_i}{\partial g_{-i}} < 0$. R_i is homogenous of degree zero: the outcome of the conflict is the same if all DTOs increase their expenditure proportionally. Spending no resources in the conflict results in controlling zero routes unless all DTOs have zero expenditure ($R_i(0, g_{-i}) = 0$ for $g_{-i} \neq 0$), and if all DTOs except i spend zero resources, i holds all routes, no matter how small its expenditure in the conflict is ($R_i(g_i, 0) = 1$ for $g_i > 0$). The marginal productivity of expenditure in the conflict is decreasing ($\frac{\partial^2 R_i}{\partial g_i^2} < 0$). As g_{-i} increases, any additional investment by DTO i is less in comparison with the size of the conflict, reducing the marginal productivity of g_i. Thus, $\frac{\partial}{\partial g_{-i}} \left(\frac{\partial R_i}{\partial g_i} \right) = \frac{\partial^2 R_i}{\partial g_{-i} \partial g_i} < 0$. I also assume that $\frac{\partial^2 R_i}{\partial g_i^2} < \frac{\partial^2 R_i}{\partial g_i \partial g_{-i}}$: consider $\frac{\partial R_i}{\partial g_i}$, the marginal productivity of g_i. It decreases both with an increase in g_i, expenditure by the same DTO, and in g_{-i}, expenditure by the other DTOs. However, if both increases are equal, the increase in g_{-i} would be spread across all other DTOs, so it is reasonable to assume that it has a milder effect on the marginal productivity. Stated mathematically, $\frac{\partial^2 R_i}{\partial g_i^2} < \frac{\partial^2 R_i}{\partial g_i \partial g_{-i}}$.[12]

In the end, since DTO i sells an amount q_i of drugs in the consumer market at a price p_c, the

[11] In order to see this, first note that $\frac{\partial^2 q}{\partial x_i \partial R_i} = x \frac{\partial^2 w}{\partial x_i \partial R_i} + \frac{\partial w}{\partial R_i}$. By using the chain rule, $\frac{\partial w}{\partial R_i} = \frac{\partial w}{\partial r_i} \frac{\partial r_i}{\partial R_i}$ and $\frac{\partial^2 w}{\partial x_i \partial R_i} = \frac{\partial^2 w}{\partial r_i^2} \frac{\partial r_i}{\partial R_i} \frac{\partial r_i}{\partial x_i} + \frac{\partial w}{\partial r_i} \frac{\partial^2 r_i}{\partial x_i \partial R_i}$. The derivatives of r_i can be readily calculated. Substituting everything in the initial expression for the cross derivative of q yields $\frac{\partial^2 q}{\partial x_i \partial R_i} = -\frac{R_i}{x_i^2} \frac{\partial^2 w}{\partial r_i^2}$, which is positive due to the decreasing marginal productivity of r_i.

[12] Although this may seem like an unusual assumption, since it does not arise in most microeconomic models, it will play a key role in determining the effect of government policy on violence.

profit it obtains is[13]:

$$\pi_i = p_c q(x_i, R_i(g_i, g_{-i}), e) - g_i - p_p x_i \tag{1}$$

DTOs interact strategically only through expenditure in the conflict: DTOs are price takers, so the amount of drugs bought in the producer market does not affect rival DTOs. Furthermore, they have no way of observing the quantity of drugs related to other DTOs, since they are part of an illegal production chain through routes they do not control[14].

3.2 Independence of productive behavior

Before turning to the various types of equilibria that may arise, I show that the aggregate productive behavior of DTOs (that is, the total amount of drugs bought at producer markets and sold at consumer markets) does not depend on the specific type of equilibrium that arises from the conflict. The crucial assumption that allows me to analyze the aggregate behavior is that function q has constant returns to scale.

Suppose that in some equilibrium the amount of routes controlled by DTO i is $\hat{R}_i$. Since the amount of drugs bought at the initial market does not affect others, i chooses the quantity that maximizes its profit, i.e.,

$$x_i^* = \underset{x_i}{\operatorname{argmax}} \left[p_c q(x_i, \hat{R}_i, e) - g_i - p_p x_i \right] \tag{2}$$

which can be solved from the following first order condition:

$$\underbrace{p_c \frac{\partial q}{\partial x_i}}_{MgBx_i} = \underbrace{p_p}_{MgCx_i} \tag{3}$$

This has the straightforward interpretation that the marginal benefit of x_i equals its marginal cost, the price at which it is bough in the producer region. The solution is a maximum due to the decreasing marginal productivity of x_i.

The previous maximization problem can be misinterpreted. I do not mean that DTOs maximize given the amount of routes they hold, since they decide the amounts of x_i and g_i simultaneously. But if the amount of drugs they buy at the producer region is not the one determined by equation (2), they can deviate and improve, and they are not at an optimum. A similar maximization problem cannot be solved to find g_i with the same generality: expenditure in the conflict has

[13] Although q has constant returns to scale, DTOs obtain a profit since they do not invest directly in routes at a fixed cost. Instead, they invest in the conflict, which has decreasing marginal productivity.

[14] If, on the other hand, DTOs had some market power, interaction through drug quantities becomes relevant. This is now the widely treated problem of a traditional Cournot oligopoly in IO.

an effect on other DTOs, which means that it is determined by the strategic interaction between DTOs, and not simply by the maximizing behavior of individual DTOs.

Since (3) holds for all cartels, and the production function is homogenous of degree one, the optimal amount of drugs bought at the initial market by each DTO is proportional to the amount of routes it holds[15]. This makes sense if we think in terms of the saturation of routes, which can be measured by $\frac{x_i}{R_i}$: different levels of saturation imply different marginal productivities of x_i, which is inconsistent with (3). A direct consequence is that the optimal amount of drugs sold by each DTO, q_i^*, is also proportional to the amount of routes it holds[16]. All this can be stated as

$$\frac{x_i^*}{x_j^*} = \frac{\hat{R}_i}{\hat{R}_j} = \frac{q_i^*}{q_j^*} \tag{4}$$

Summing over all DTOs yields $\frac{\sum_i x_i^*}{x_j^*} = \frac{\sum_i \hat{R}_i}{\hat{R}_j} = \frac{\sum_i q_i^*}{q_j^*} \implies \frac{X}{x_j^*} = \frac{1}{\hat{R}_j} = \frac{Q}{q_j^*}$. Therefore, the aggregate amount of drugs bought X is $\frac{x_j^*}{\hat{R}_j}$, the unique proportion between drugs bought and routes held determined by equation (3), and it is independent of the way routes are distributed between DTOs. It can be found by solving the following first order condition for X:

$$p_c \frac{\partial q(X,1,e)}{\partial x} = p_p \tag{5}$$

The aggregate supply of drugs is $Q = \sum_i q(\hat{R}_i X, \hat{R}_i, e) = \sum_i \hat{R}_i q(X,1,e) = q(X,1,e)$, and it is therefore also independent of the distribution of routes among DTOs.

Proposition 1. *The aggregate productive behavior (i.e., X and Q) is independent of the strategic interaction between DTOs.*

Proposition 1 is a very general result. It is a consequence of the production technology having constant returns to scale, and of the fact that the conflict is a zero-sum game, at the end of which the amount of routes controlled by all DTOs is the same. It is true regardless of how the conflict over routes is solved, and in particular, it does not change if routes are split asymmetrically. For instance, if some DTOs are more powerful than others, as long as they are equally efficient in terms of transporting drugs given an amount of routes, the supply of drugs will not change.

It is natural to ask what share of the total amount of drugs is bought and sold by each DTO. I had already shown that $X = \frac{x_j^*}{\hat{R}_j}$, and $q_i = q(x_i, \hat{R}_i, e) = \hat{R}_i q(X,1,e)$.

[15] For two different cartels i and j, $\frac{\partial q(x_i^*, R_i^*, e)}{\partial x_i} = \frac{\partial q(x_j^*, R_j^*, e)}{\partial x_j}$. Since q is homogeneous of degree one in its first two arguments, its derivative with respect to x_i is homogeneous of degree zero, so $\frac{\partial q(x_i^*/R_i^*, 1, e)}{\partial x_i} = \frac{\partial q(x_j^*/R_j^*, 1, e)}{\partial x_j}$, and since this derivative is strictly decreasing, $\frac{x_i^*}{R_i^*} = \frac{x_i^*}{R_i^*}$.

[16] $q_i^* = q(x_i, R_i, e) = q(\frac{x_i^*}{x_j^*} x_j^*, \frac{R_i}{R_j} R_j, e) = \frac{R_i}{R_j} q_j^*$.

Proposition 2. *The share of drugs bought and sold by each DTO is equal to the fraction of routes it controls:*

$$x_i = R_i X \qquad q_i = R_i Q \tag{6}$$

I can now find a unique expression for the *aggregate productive profit*, which I denote by π^A. It is the sum of the profits obtained by all DTOs, without taking into account their expenditure in the conflict. Thus, it takes into account productive behavior but not military behavior:

$$\pi^A = p_c Q - p_p X \tag{7}$$

Proposition 2 allows me to obtain a new expression for DTO i's profits. It invests g_i in the conflict, buys an amount of drugs $x_i = R(g_i, g_{-i})X$ and sells an amount $q_i = R(g_i, g_{-i})Q$. Substituting these amounts in (1) results in $\pi_i = p_c q_i - p_p x_i - g_i = (p_c Q - p_p X)R(g_i, g_{-i}) - g_i$, which can be written in terms of the aggregate productive profit:

Proposition 3. *The trafficking industry can be restated as*

$$\pi_i = \pi^A R(g_i, g_{-i}) - g_i \tag{8}$$

In other words, DTOs fight over the aggregate productive profit, $\pi^A = p_c Q - p_p X$, and they spend resources in the conflict in order to obtain a share of the prize. In the end, they end up receiving a fraction of π^A that is equal to the fraction of routes they control.

The effect of enforcement on the aggregate amount of drugs bought can be found from the total differential of (5) when the only exogenous variable that changes is e:

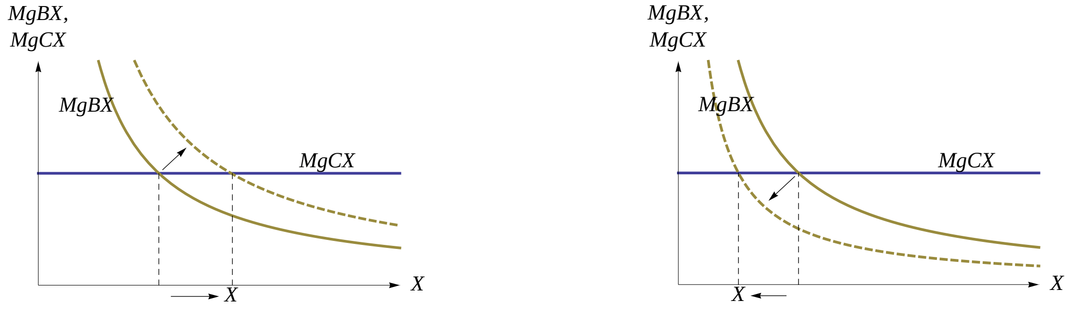
$$\left[\frac{dp_c}{dQ}\left(\frac{\partial q}{\partial X}\right)^2 + p_c\frac{\partial^2 q}{\partial X^2}\right] dX = -\left[\frac{dp_c}{dQ}\frac{\partial q}{\partial e}\frac{\partial q}{\partial X} + p_c\frac{\partial^2 q}{\partial X \partial e}\right] de \tag{9}$$

All derivatives of q must be evaluated at $(X, 1, e)$. I exploit the fact that, since this is the aggregate production function, the amount of routes in the second argument of q is the total amount of routes, so it is constant. The term $\frac{dp_c}{dQ}$ can be written in terms of elasticities. After some simplifications, this leads to

$$\frac{\partial X}{\partial e} = \left(\frac{-1}{\frac{1}{Q\epsilon_c}\left(\frac{\partial q}{\partial X}\right)^2 + \frac{\partial^2 q}{\partial X^2}}\right)\left[\overbrace{\frac{1}{Q\epsilon_c}\frac{\partial q}{\partial X}\frac{\partial q}{\partial e}}^{\text{(a)}} + \overbrace{\frac{\partial^2 q}{\partial X \partial e}}^{\text{(b)}}\right] \tag{10}$$

The term in parentheses is positive, so the sign is determined by the term in square brackets.

Two mechanisms are at work, and they can be clearly seen in equation (10). First, enforcement decreases the amount of drugs that reach the consumer market, if the amount of drugs bought



(a) Effect from productivity $\frac{\partial q}{\partial e}$.

(b) Effect from marginal productivity $\frac{\partial^2 q}{\partial e \partial X}$.

Figure 1: The two effects of enforcement on the amount of drugs bought.

from the producer market were held fixed. This leads to an increase in prices, which increases the marginal benefit of drugs and encourages DTOs to take more drugs from the producer to the consumer region (see figure 1a). This mechanism is represented by term (a). Second, enforcement reduces the marginal productivity of X, and thus its marginal benefit. In response to this, DTOs buy a smaller amount of drugs so that marginal benefits and costs of X are again equal (see figure 1b). The term denoted with (b) represents this mechanism. Enforcement thus increases or decreases the amount of drugs bought depending on which effect is larger. This depends, in particular, on whether demand is inelastic enough that effect (b) is larger. Note that if the trafficking region has a small share of the supply, $\epsilon_c = \infty$, and mechanism (b) has no effect on X.

The chain rule can be used to find the effect on the supply of drugs:

$$\frac{\partial Q}{\partial e} = \frac{\partial q}{\partial x}\frac{\partial X}{\partial e} + \overbrace{\frac{\partial q}{\partial e}}^{\text{(c)}} \tag{11}$$

Since $\frac{\partial q}{\partial x} > 0$, mechanisms (a) and (b) are still at work here, but now there is a third mechanism (c): For a fixed amount of drugs bought in the producer market, the amount of drugs that reach the producer market decreases with enforcement due to the decreased productivity of X. It would thus seem that supply may increase or decrease with enforcement, again depending on the elasticity of demand. But increased supply means a decrease in prices. In that case mechanism (b), the only one through which supply increases, would work the other way around, also decreasing supply. Thus, an increase in supply is a contradiction: it is inconsistent with the increase in prices that would cause it.

Formally, this can be seen by substituting (10) in (11), which yields

$$\frac{\partial Q^e}{\partial e} = \frac{\frac{\partial^2 q}{\partial X^2}\frac{\partial q}{\partial e} - \frac{\partial q}{\partial X}\frac{\partial^2 q}{\partial X \partial e}}{\frac{1}{Q\epsilon_c}\left(\frac{\partial q}{\partial X}\right)^2 + \frac{\partial^2 q}{\partial X^2}} \tag{12}$$

By looking at the individual derivatives, it is clear that enforcement reduces supply.

Proposition 4. *The comparative statics on the amount of drugs supplied to the consumer region is:*

- $\dfrac{\partial Q}{\partial e} < 0$*: Increasing enforcement reduces supply of drugs.*
- $\dfrac{\partial Q}{\partial n} = 0$*: The number of DTOs has no effect on the supply of drugs*

In terms of public policy, this proposition means that actions taken by the government to affect the conflict between cartels may affect violence, but they have no effect on the amount of drugs reaching final consumer markets. Therefore, actions taken by the government should aim against productive behavior (i.e., enforcement) if they are to reduce the supply of drugs in consumer markets.

3.3 Stage-game Nash equilibrium (SGNE)

The conflict over routes in illegal drug markets has already been treated as a one-period game, in which DTOs are in a Nash equilibrium. Nevetheless, solving the stage game will serve as a benchmark for comparison, and it will be useful in order to prove my main results, which are related to repeated interaction.

DTOs do not care about the future if they interact for a single period. Thus, they maximize their instantaneous profit given expenditure by all other DTOs. Consider an individual DTO that observes others' behavior (i.e., their expenditure in the conflict g_{-i}), and, based on that, maximizes its profit. Interpreting the conflict as in proposition (3), the problem it faces is:

$$\max_{g_i} \pi_i = \pi^A R(g_i, g_{-i}) - g_i \tag{13}$$

which leads to the following first order condition:

$$\underbrace{\pi^A \frac{\partial R}{\partial g_i}}_{MgBg_i} = \underbrace{1}_{MgCg_i} \tag{14}$$

Condition (14) is easy to interpret: it states that the marginal benefit of investment in the conflict must equal its marginal cost (one). It seems that the second order condition is fulfilled, since $\frac{\partial^2 R}{\partial g_i^2} < 0$. However, I am maximizing in two steps, (2) and (13), without taking into account that the maximization is actually done simultaneously. Thus, I should check the second order conditions for the complete problem faced by the DTO, without isolating productive and military behavior:

$$\max_{(g_i, x_i)} \pi_i = p_c q(x_i, R_i(g_i, g_{-i}), e) - g_i - p_p x_i \tag{15}$$

The first order conditions for this problem are (3) and

$$\underbrace{p_c \frac{\partial q}{\partial R_i} \frac{\partial R_i}{\partial g_i}}_{MgBg_i} = \underbrace{1}_{MgCg_i} \tag{16}$$

I assume that $\frac{p_c}{p_p}$ is large enough that there is an interior solution[17], since otherwise the illegal drug market would not even exist. Concavity of both q_i and R_i ensures that any solution to both first order conditions is indeed a maximum[18].

It is easier to use (14), since is simpler than (16), but it still remains unclear whether both solve equivalent problems, which would allow me to conclude that (14) is indeed a maximum. This can be solved from homogeneity of degree one of q and Euler's homogeneous function theorem. Homogeneity of degree one means that derivatives are homogeneous of degree zero, so $\frac{\partial q}{\partial R_i}$ is the same if it is evaluated for an individual DTO, at (x_i, R_i, e), and for the aggregate industry, at $(X, 1, e)$. Euler's theorem means that $Q = X\frac{\partial q(X,1,e)}{\partial X} + \frac{\partial q(X,1,e)}{\partial R}$, and from (5), $p_c \frac{\partial q}{\partial R_i} = p_c Q - p_p X = \pi^A$, meaning that both first order conditions are equivalent.

First order conditions (14), one for each DTO, give the best-response functions for g_i in terms of the quantities g_{-i} chosen by all other DTOs. The Nash equilibrium of the stage game occurs when the first order conditions are fulfilled simultaneously for all DTOs, i.e., when all best-response functions are consistent. I make the additional assumption that all DTOs are equal. I had treated them so far as equally efficient in their productive behavior. Now I also treat them equally in the conflict for routes, meaning that functions R_i and R_j are equal for any two DTOs i and j. Thus, no DTO has an advantage in the conflict for routes.

The symmetry of the problem means that $g_i = g_j = g^N, \quad x_i = x_j = x^N \quad \forall i, j \in I$ (the superscript refers to the solution being the one-period Nash equilibrium), and $g_{-i} = (n-1)g^N$. Furthermore, every DTO controls an equal amount of routes $R_i = \frac{1}{n}$. Proposition 4 already stated the comparative statics on the amount of drugs taken to the consumer region. In order to find the effect of enforcement on the total amount of violence, I analyze the effect on the individual level of violence. Since the number of DTOs is fixed, any change in individual expenditure induces a change in the level of violence in the same direction.

Individual expenditure is determined by first order condition (14). Note that enforcement has

[17] There cannot be a corner solution with $g_i = 0$ and $x_i > 0$, since the marginal productivity of expenditure in the conflict tends to infinity if all DTOs spend zero resources in the conflict.

[18] The second-order conditions are $\frac{\partial^2 \pi_i}{\partial x_i^2} = p_c \frac{\partial^2 q_i}{\partial x_i^2} < 0$, $\frac{\partial^2 \pi_i}{\partial g_i^2} = p_c \left[\frac{\partial q_i}{\partial R_i} \frac{\partial^2 R_i}{\partial g_i^2} + \frac{\partial^2 q_i}{\partial R_i^2} \left(\frac{\partial R_i}{\partial g_i} \right)^2 \right] < 0$, and $\frac{\partial^2 \pi_i}{\partial x_i^2} \frac{\partial^2 \pi_i}{\partial g_i^2} - \left(\frac{\partial^2 \pi_i}{\partial x_i \partial g_i} \right)^2 = p_c^2 \left[\frac{\partial^2 q_i}{\partial x_i^2} \frac{\partial q_i}{\partial R_i} \frac{\partial^2 R_i}{\partial g_i^2} + \left(\frac{\partial R_i}{\partial g_i} \right)^2 \left(\frac{\partial^2 q_i}{\partial x_i^2} \frac{\partial^2 q_i}{\partial R_i^2} - \left(\frac{\partial^2 q_i}{\partial x_i \partial R_i} \right)^2 \right) \right] > 0$. Strict concavity of R_i and the concavity of q_i ensure that all three conditions are fulfilled.

no direct effect on $\frac{\partial R_i}{\partial g_i}$, whereas g^N has no direct effect on π^A. This, and the implicit function theorem, lead to the following expression for the effect of enforcement on expenditure in the conflict:

$$\frac{\partial g^N}{\partial e} = -\frac{\frac{\partial \pi^A}{\partial e}}{\frac{\partial}{\partial g^N}\left(\frac{\partial R}{\partial g_i}\right)} \tag{17}$$

The denominator is negative: $\frac{\partial^2 R}{\partial g^N \partial g_i} = \frac{\partial^2 R}{\partial g_i^2} + (n-1)\frac{\partial^2 R}{\partial g_{-i}\partial g_i} < 0$. Thus, the sign of the effect of enforcement is the same as the sign of the effect on π^A, the prize being fought over. This is easy to interpret: the conflict intensifies as the stakes become larger.

It is not clear whether enforcement increases or decreases productive profit $\pi^A = p_c Q - p_p X$. The question is whether $\frac{\partial \pi^A}{\partial e}$ is positive or negative. It can be expanded in terms of costs and revenues: $\frac{\partial \pi^A}{\partial e} = \frac{\partial p_c}{\partial Q}\frac{\partial Q}{\partial e}Q + p_c \frac{\partial Q}{\partial e} - p_p \frac{\partial X}{\partial e}$. Rewriting $\frac{\partial p_c}{\partial Q}$ in terms of the elasticity of demand leads to

$$\frac{\partial \pi^A}{\partial e} = p_c\left(1 + \frac{1}{\epsilon_c}\right)\frac{\partial Q}{\partial e} - p_p \frac{\partial X}{\partial e} \tag{18}$$

Substituting $\frac{\partial Q}{\partial e}$ and $\frac{\partial X}{\partial e}$ from (10) and (12) allows me to find a threshold for the elasticity of demand such that productive profit increases if $\epsilon_c > \hat{\epsilon}_c$: enforcement causes an increase in prices large enough that aggregate productive profit increases. Thus, the prize DTOs fight for increases, increasing the marginal benefit of expenditure in the conflict. The expression for this threshold is:

$$\hat{\epsilon}_c = -1 - \underbrace{\frac{\left(\frac{\partial q}{\partial X}\right)^2}{\frac{\partial q}{\partial e}\frac{\partial^2 q}{\partial X^2}}}_{\text{(a)}} \underbrace{\left(\frac{\frac{\partial q}{\partial e}}{Q} - \frac{\frac{\partial^2 q}{\partial X \partial e}}{\frac{\partial q}{\partial X}}\right)}_{\text{(b)}} \tag{19}$$

This expression is an improvement on the threshold of -1 that determines whether revenues (equal to the market size, $p_c Q$) increase or decrease in response to reduced supply. The analysis by (Becker et al., 2006) that leads to this threshold measured the decrease in welfare in consumer nations under competitive markets, where market size is a good measure of the harm caused by illegal drugs. However, subsequent works have taken this threshold out of context, and they have concluded that it also determines the range of elasticities over which supply reduction increases violence. This is a mistake, since it does not take into account DTOs' costs, which may increase or decrease in response to supply, with opposing effects on violence. Equation (19) is thus the threshold of -1, plus a correction that tells if it is easier or harder for enforcement to increase violence than with the original threshold. Term (a) is negative, so the correction's effect is determined by the sign of term (b), which can be written as $\frac{\partial \log q}{\partial e} - \frac{\partial \log \frac{\partial q}{\partial X}}{\partial e}$: whether enforcement has a larger effect on *productivity* or on *marginal productivity* of drugs. I will now explain why this difference determines the sign of the correction.

Since the correction comes from costs $p_p X$, and p_p is fixed, we are now interested in the effect of enforcement on X. I already analyzed this in section 3.2: enforcement has two effects on X. Due to its effect on productivity ($\frac{\partial \log q}{\partial e}$), X increases (figure 1a), as well as costs, which means that it is possible for profit to decrease while revenue increases. Hence, the effect of enforcement on productivity makes it more difficult for profit to increase, which explains why it makes the threshold higher (more restrictive)[19]. On the other hand, due to enforcement's effect on marginal productivity ($\frac{\partial \log \frac{\partial q}{\partial X}}{\partial e}$), X decreases (figure 1b), which makes it easier for profit to increase, and the threshold is therefore lower (less restrictive). The term $\left(\frac{\partial \log q}{\partial e} - \frac{\partial \log \frac{\partial q}{\partial X}}{\partial e}\right)$ thus captures whether costs increase or decrease with enforcement, which depends on which is larger: the effect through productivity, or the effect through marginal productivity.

If costs are low the correction should be smaller. This is captured by term (a): from first order condition (5), $\frac{\partial q}{\partial X} = \frac{p_p}{p_c}$. Thus, as prices in the producer market become lower in comparison with the prices at which DTOs sell drugs, costs take a smaller share of revenues and the correction becomes less important. Enforcement reduces supply, increasing prices. Thus, enforcement makes the correction to the threshold less important.

I analyzed the threshold as far as it is possible without specifying a particular functional form for q. By setting some particular functional form, it is possible to determine whether the threshold is more or less restrictive than -1. I undertake this analysis in section 3.5. The main result is that for the production technology of an illegal drug market the effect of enforcement on marginal productivity is larger than the effect on productivity. Thus, $\hat{\epsilon}_c < -1$, and demand does not necessarily have to be inelastic for enforcement to increase violence.

In order to find the effect of the number of cartels on violence in a SGNE, I analyze first order condition (14), $\pi^A \frac{\partial R}{\partial g_i} = 1$. Since π^A does not depend on the number of cartels, $\frac{\partial R_i}{\partial g_i}$ cannot depend on it either, which means that its derivative with respect to n must be zero:

$$\frac{\partial^2 R}{\partial n \partial g_i} = \left[\frac{\partial^2 R}{\partial g_i^2} + (n-1)\frac{\partial^2 R}{\partial g_i \partial g_{-i}}\right]\frac{\partial g^N}{\partial n} + g^N \frac{\partial^2 R}{\partial g_i \partial g_{-i}} = 0 \tag{20}$$

From this expression, $\frac{\partial g^N}{\partial n}$ can now be isolated:

$$\frac{\partial g^N}{\partial n} = -g^N \frac{\partial^2 R}{\partial g_i \partial g_{-i}}\left[\frac{\partial^2 R}{\partial g_i^2} + (n-1)\frac{\partial^2 R}{\partial g_i \partial g_{-i}}\right]^{-1} < 0 \tag{21}$$

The sign of this expression is negative, as can be seen from the individual derivatives.

In order to find the comparative statics on aggregate violence G, I use the fact that $\frac{\partial G^N}{\partial n} =$

[19] Strange as it may seem, $\hat{\epsilon}_c$ can be positive: the effect of enforcement on productivity would be so large that even with perfectly inelastic demand enforcement would decrease profits.

$g^N + n\frac{\partial g^N}{\partial n}$ to obtain

$$\frac{\partial G^N}{\partial n} = g^N \left[\frac{\partial^2 R}{\partial g_i^2} - \frac{\partial^2 R}{\partial g_i \partial g_{-i}}\right] \left[\frac{\partial^2 R}{\partial g_i^2} + (n-1)\frac{\partial^2 R}{\partial g_i \partial g_{-i}}\right]^{-1} > 0 \tag{22}$$

which is positive.

The intuition behind this result is that a greater number of DTOs decreases the fraction of total routes each one of them holds, increasing the marginal productivity of routes. This, in turn, leads to each DTO spending more resources in the conflict. The outcome is a rat race since total expenditure increases, but they all control the same total amount of routes. The main assumption behind this result is therefore the fact that resources spent in the conflict have diminishing returns to scale.

The following proposition summarizes the results regarding the SGNE:

Proposition 5. *Under a symmetric stage-game Nash equilibrium, the comparative statics on the level of violence in the region is as follows:*

- *If $\epsilon_c < \hat{\epsilon}_c$, then $\frac{\partial G^N}{\partial e} < 0$: If demand is sufficiently elastic, enforcement reduces the level of violence.*
- *If $\epsilon_c > \hat{\epsilon}_c$, then $\frac{\partial G^N}{\partial e} > 0$: If demand is sufficiently inelastic, enforcement increases the level of violence.*
- *$\frac{\partial G^N}{\partial n} > 0$: An increase in the number of DTOs increases the level of violence.*

The main novelty introduced by proposition 5 is that it refines the elasticity threshold $\hat{\epsilon}_c = -1$ that has been used traditionally in the literature.

3.4 Repeated interaction and collusion

I will now consider the same situation described before, but being repeated for multiple periods. The total profits obtained by a DTO are the discounted sum of the profit obtained in each of the periods, namely

$$\Pi_i = \sum_{t=0}^{\infty} \beta^t \pi_{i,t} \tag{23}$$

where $\pi_{i,t}$ is the profit obtained by DTO i in period t, and $\beta \in (0,1)$ is the discount factor. Note that I will distinguish between the one-period profit, denoted by π, and the discounted sum of all profits, denoted by Π. The discount factor depends on two different elements: the monetary discount factor related to the interest rate, which I call δ, and the probability p that the current leader of the DTO will still be in charge in the next period. The discount factor is then $\beta = \delta p$. The probability depends on the government's actions, since policies aimed at capturing or killing leaders decrease the probability that they will be standing during the next period. This means that DTO leaders are selfish, since they do not value the future of their organization after they are captured or killed. Hence, they care less about the future if they believe that there is a significant probability of being killed soon.

Repeated interaction makes many more strategies available to any DTO, since they can now respond to the actions taken by other DTOs in previous periods. The baseline strategy is simply repeating the non-cooperative Nash equilibrium from the stage game perpetually, which results in each DTO obtaining a profit $\Pi^N = \dfrac{\pi^N}{1-\beta}$. They can also choose to agree to peace treaties, or at least milder fights, which can be enforced by future punishment. I will now analyze some strategies that allow them to decrease expenditure in the conflict, allowing higher profits for all DTOs[20].

3.4.1 Peaceful equilibrium

In an ideal collusive treaty, DTOs agree to split routes evenly between them, without any expenditure in the conflict. Each DTO controls $\frac{1}{n}$ routes, and, from proposition 3, each DTO obtains profits

$$\pi^c(0) = \frac{1}{n}\pi^A \tag{24}$$

where $\pi^c(0)$ stands for the profit obtained under collusion with zero expenditure in the conflict.

Let us see whether any particular DTO has incentives to deviate from this collusive equilibrium. If some DTO betrays by increasing its expenditure, in the next period all other DTOs would punish it. One possible punishment is returning to the SGNE, after which no single DTO could deviate to its advantage. This punishment strategy, called *Nash reversion*, has been widely studied in repeated games, and it is used in industrial organization as a punishment for firms that deviate from collusion in oligopoly (see Motta, 2004). But if players could find a harsher punishment than Nash reversion, they could create greater incentives to stay at the collusive equilibrium (Abreu, 1983). The harshest punishment is making sure that any DTO that betrays receives zero profits from that moment on, since any lower profits would encourage the betrayer to exit from the industry. I refer to punishment with zero profits for the traitor as *optimal punishment*.

In traditional oligopolies firms have difficulties enforcing the treaties they agree to. First, they only observe prices, so they can only see if some firm deviates, but they cannot determine which

[20] DTOs could agree to follow oligopolist treaties, in which they reduce supply in order to increase prices. However, I will not consider this type of treaties since I assume that DTOs are price takers.

firm deviated. Additionally, the method they use to punish are price wars that do not have a specific firm as target, so all firms are equally punished, regardless of whether they deviated or not. This is no longer the case with DTOs: they are violent organizations that will readily attack any betrayer, and it is easy to see which DTO deviated from the collusive equilibrium. Thus, I assume that a punishment that induces zero profits on the betrayer can be achieved if all DTOs join their forces against it[21]. Since I model the conflict as a non-directed contest, in which expenditure in the conflict does not have a specific DTO as target, this punishment does not fit explicitly in my model, but I nevertheless assume that DTOs know that other DTOs can use optimal punishment on them[22].

If all DTOs invest zero resources in the conflict, they do not have the means to defend their routes. Thus, any single DTO can invest an infinitesimal amount η in the conflict, and it will be able to control all routes. For a single period, the traitor takes all the aggregate productive profit, $\pi^t(0) = \pi^A - \eta = n\pi^c(0)$. From the next period on it will receive zero profits due to others' punishment. On the other hand, if it complies with the treaty, it obtains profits $\frac{1}{1-\beta}\pi^c(0)$. The punishment is strong enough to dissuade betrayal if $\frac{1}{1-\beta}\pi^c(0) \geq \pi^t(0)$. The last expression is the *incentive constraint* (IC) that must be fulfilled for collusion to hold. Isolating β from the IC yields the following result:

Proposition 6. *A peaceful collusive equilibrium can be sustained if and only if* $\beta \geq \dfrac{n-1}{n}$.

Proposition 6 means that under the right circumstances cartels can coexist without any violence. This depends on two conditions: there must be a low number of cartels, and they must place a high value on future earnings. Even though an individual DTO could betray and seize all routes for one period at a negligible cost, meaning huge benefits in the short run, this would also mean reducing its profits in the long run. If DTOs are sufficiently fearful of the future reduction in their profits, they will not want to betray, no matter how easy it is for them to take all the routes. On the other hand, if there are more than a few cartels, by seizing all routes they would obtain a great increase in profits, which would require a very high discount factor for a totally peaceful equilibrium to exist.

It may seem strange that enforcement has no effect on whether peace can be sustained or

[21] This would seem to violate the assumption that the number of DTOs is fixed, if the punished betrayer then exited the industry. However, this will never happen: it is optimal for DTOs not to betray, and therefore not being punished.

[22] An important condition for the punishment strategy is that it should also be subgame perfect, or in layman's terms, it should be a credible threat. This means that if one DTO betrays, other DTOs should not have incentives to deviate from punishing the traitor. The way to achieve this is to treat any DTO that does not cooperate in the punishment as a traitor, thus making it receive zero profits from that moment on. I assume that punishing is indeed subgame perfect, and as a consequence the collusive equilibrium is a subgame perfect equilibrium.

not. The reason behind this is that enforcement reduces both the profits of colluding and treason through productive profits. In a peaceful equilibrium betrayers obtain the full productive profit, whereas colluders obtain a constant fraction $\frac{1}{n}$. Thus, enforcement has the same effect on betraying and colluding, and it makes the existence of peaceful equilibrium neither easier nor harder to sustain.

3.4.2 Collusive equilibrium with violence

If peace cannot be sustained (if $\beta < \frac{n-1}{n}$), one would not expect DTOs to wage all-out war, as in the SGNE. They can still agree on spending $g_r < g^N$ on the conflict, after which they end up controlling the same amount $R^c = \frac{1}{n}$ of routes as in the SGNE but with a higher profit[23]. The purpose of $g_r > 0$ is to make betrayal more costly than with zero expenditure, when an infinitesimal expenditure was enough to grab all routes. Thus, punishment and g_r work together as deterrents against collusion. I will therefore call g_r *deterrent expenditure*. The profit obtained by each DTO is then

$$\pi^c(g_r) = \pi^A R^c - g_r = \frac{1}{n}\pi^A - g_r \tag{25}$$

The first order condition is again (3), and productive behavior is the same as in the stage game. The only difference between this collusive agreement and the Nash equilibrium is the lower level of investment in the conflict by each DTO. Comparing the profit from colluding from the profit at the SGNE yields $\pi^c(g_r) = \pi^N + g^N - g_r$.

What determines the amount g_r spent in the conflict by each DTO? They would benefit if they could all spend a low amount, since the collusion treaty means that regardless of the amount spent they all hold the same fraction of routes. However, if that amount is too low, any particular DTO would be able to break the treaty and invest a larger amount of resources in order to attack other DTOs and grab a larger fraction of the routes, thereby increasing its profit. Thus, g_r cannot be arbitrarily low: it will be as low as possible while still working as a deterrent against betraying.

Since the benefit of deviating from the cooperative strategy only lasts for one period, the traitor DTO would want to take as much profit as it can for that single period, i.e., the optimal one-period behavior given that all other DTOs spend g_r. Thus, by interpreting the conflict as in proposition 3, the profit obtained by the traitor i when it betrays is:

$$\pi^t(g_r) = \max_{g_i} \left[\pi^A R_i(g_i, (n-1)g_r) - g_i\right] \tag{26}$$

[23] This was not possible in the stage game because a lower level of expenditure by all other DTOs meant an increase in the marginal utility of expenditure for every DTO (since $\frac{\partial^2 q}{\partial g \partial g_{-i}} < 0$), implying an incentive to deviate unilaterally and increase expenditure.

The first order condition is the same as for the SGNE, but with expenditure by other DTOs evaluated at $g_{-1} = ng_r$. As in the SGNE, I stated the problem as two maximizations over different variables, but the DTO solves a joint maximization problem over both variables. Thus, I must check the optimality of the joint maximization problem:

$$\pi^t(g_r) = \max_{x_i, g_i} \left[p_c q(x_i, R_i(g_i, (n-1)g_r), e) - g_i - p_p x_i \right] \tag{27}$$

The second order conditions are the same as for the problem in the SGNE, which implies that it leads to a maximum.

Solving (26) results in an optimal expenditure in the conflict g^t, that determines the optimal amount of routes $R^t = R(g^t, (n-1)g_r)$. The betrayer's profit is then

$$\pi^t(g_r) = \pi^A R^t - g^t \tag{28}$$

The punishment strategy used, either Nash reversion or optimal punishment, determines the lowest level g_r that sustains a collusive equilibrium. Since optimal punishment maximizes incentives against betraying, it allows g_r to be as low as it can be. My aim will be to analyze what happens in that case. I will also look at what happens with Nash reversion, but only because it will be useful in proving some results regarding optimal punishment.

Let π^p be the profit when being punished ($\pi^p = \pi^N$ for Nash reversion, $\pi^p = 0$ for optimal punishment). The total profits of the traitor are $\Pi^t = \pi^t + \frac{\beta}{1-\beta}\pi^p$, and the total profits if it does not deviate from collusion are $\Pi^c = \frac{1}{1-\beta}\pi^c$. This means that no DTO would have an incentive to deviate from the cooperative equilibrium if $\Pi^c(g_r) \geq \Pi^t(g_r)$, namely, if[24]

$$\frac{1}{1-\beta}\pi^c(g_r) \geq \pi^t(g_r) + \frac{\beta}{1-\beta}\pi^p \quad \Longleftrightarrow \quad \pi^c(g_r) \geq (1-\beta)\pi^t(g_r) + \beta\pi^p \tag{29}$$

If there exists some reserve level of expenditure in the conflict such that the new IC (29) is fulfilled, the collusive equilibrium can be sustained. Such level *always* exists, even when the punishment is Nash reversion: by setting $g_r = g^N$, $\pi^c(g_r)$ becomes π^N, since all DTOs will be cooperating with the conflict expenditure corresponding to the SGNE. The one-period optimal response to cooperation with $g_r = g^N$ is a level of investment g^N, so $\pi^t(g_N)$ is π^N as well. The IC is thus fulfilled with equality at $g_r = g^N$.

From (25), $\frac{\partial \pi^c}{\partial g_r} = -1$. On the other hand, from the envelope theorem on (26), $\frac{\partial \pi^t(g^N)}{\partial g_r} = \pi^A \frac{\partial R}{\partial g_r}\Big|_{g_r = g^N} - 1$. But $\frac{\partial R}{\partial g_r} = 0$, so $\frac{\partial \pi^t(g^N)}{\partial g_r} = -1$. This means that at $g_r = g^N$ the derivative of

[24] IC (29) can be written in terms of the discounted value of the profits (left side), or in terms of the one-period profit that should be earned as a perpetuity (right side). The difference is a factor of $(1-\beta)$. Working with one-period profits simplifies the algebra.

the left hand side of the IC is -1, which is lower than the derivative of the right hand side, $-(1-\beta)$. Hence, by setting g_r infinitesimally below g^N, the IC holds strictly, and in conclusion there exists some level $g_r < g^N$ for which the IC holds[25]. With optimal punishment, the profit from betraying is lower than with Nash reversion. Thus, any level of g_r that fulfills the IC with Nash reversion also fulfills it with optimal punishment, ensuring the existence of some $g_r < g^N$ that allows collusion with optimal punishment.

Proposition 7. *A collusive equilibrium with less violence than the SGNE always exists.*

Proposition 7 means that rational DTOs always collude, resulting in lower levels of violence than in the SGNE. Therefore, previous analyses missed an important part of the behavior of DTOs: they never engage with the level of violence predicted by one-period models. Instead, it is to their benefit to spend less in the conflict.

I will now analyze what determines the precise level of violence that allows collusion. In order to maximize profits, DTOs spend the minimum amount that ensures that the IC is fulfilled. I denote these amounts with $g^{c,N}$ for Nash reversion, and g^c for optimal punishment. They are defined by

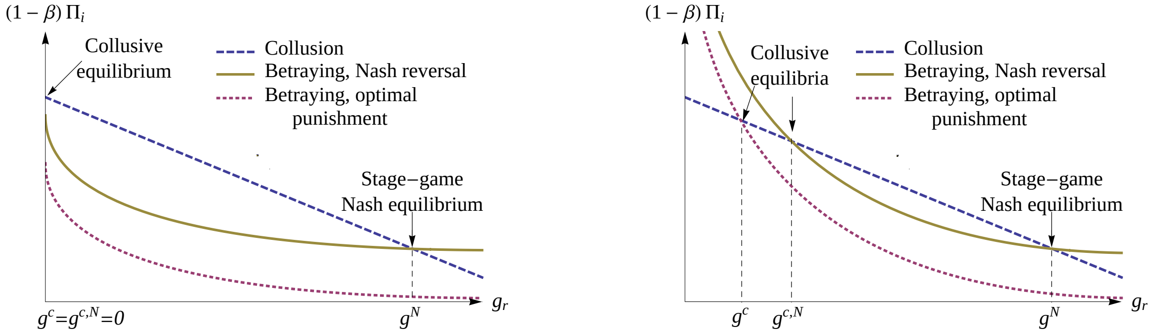
$$\pi^c(g^{c,N}) = (1-\beta)\pi^t(g^{c,N}) + \beta\pi^N \qquad \pi^c(g^c) = (1-\beta)\pi^t(g^c) \tag{30}$$

When compared with the stage game, this equation replaces second order condition (16): it determines expenditure in the conflict by each DTO, and the aggregate level of violence, in terms of the incentives that must be fulfilled in order for collusion to hold. It differs greatly from the first order condition that determines the conflict in one-period models, by making the marginal cost and benefit of expenditure in the conflict equal.

The previous analysis can be understood more easily by looking at it graphically. Let us first consider the profit from betraying if the punishment is Nash reversion. Figure 2 shows how two different cases can arise. If the discount factor is high, meaning that DTOs are very forward-looking, $\Pi^t(g_r)$ crosses $\Pi^c(g_r)$ only once, at $g_r = g^N$: even with zero investment in the conflict DTOs would prefer not to betray, since returning to the SGNE would mean a harm greater than the potential benefit from betraying. This allows the existence of a peaceful equilibrium. But if the discount factor is low, there might be a second crossing, which determines the level of investment in the conflict by each DTO, since it is the minimum value for which the IC is fulfilled[26].

[25] This is a particular case of a general proof in Mas-Colell et al. (1995), chapter 12 appendix A, that states that any SGNE can be improved by using Nash reversion strategies.

[26] It would seem that a third possibility exists. If the derivative of the right side of the IC at the SGNE were lower than the derivative of the left side, g^N would be the lowest level for which the IC is fulfilled, i.e., $\bar{g_r} = g^N$. However, the derivative of the left side is greater than the derivative of the right side, regardless of the functional forms used.



(a) High β: The future is important enough that with zero investment in the conflict the collusive equilibrium can be sustained.

(b) Low β: betraying is relatively more profitable, so DTOs would betray if they were in a peaceful equilibrium. However, there still exist levels of investment $g_r > g^N$ that allow collusion to exist.

Figure 2: Determination of the collusive equilibrium for a fixed number of DTOs.

If DTOs use optimal punishment, the profits from betraying are $(1-\beta)\pi^N$ lower than with Nash reversion. Figure 2 shows these profits as the curve with Nash reversion displaced downwards. A peaceful equilibrium and an equilibrium with some violence can also arise, depending on the discount factor. From the figure it is clear that when there is violence the amount each DTO spends in the conflict with optimal punishment, g^c, is lower than the amount spent with Nash reversion, $g^{c,r}$. There remains the possibility that for intermediate values of β a peaceful equilibrium exists with optimal punishment but not with Nash reversion.

The quantity of drugs that reach the final market is still the same, from proposition 1. Therefore, the comparative statics is the same as with the SGNE. It is also interesting to find the comparative statics with respect to the discount factor. Since the aggregate productive profit is the result of a single-period maximization, the discount factor does not affect it. This justifies the following proposition:

Proposition 8. *Under a collusive equilibrium, the comparative statics on the total amount of drugs taken to the consumer market is as follows:*

- $\dfrac{\partial Q^c}{\partial e} < 0$*: Greater expenditure by the government in enforcement reduces supply.*
- $\dfrac{\partial Q^c}{\partial n} = 0$*: The number of DTOs has no effect on supply.*

Note that nothing precludes the profit of betrayal from being concave, which would only mean that it would be easier for the peaceful equilibrium to exist. However, one would expect it to be convex at least for a very low level of g_r, since the initial reserve expenditure in the conflict has a very strong impact on whether it would be beneficial for DTOs to betray.

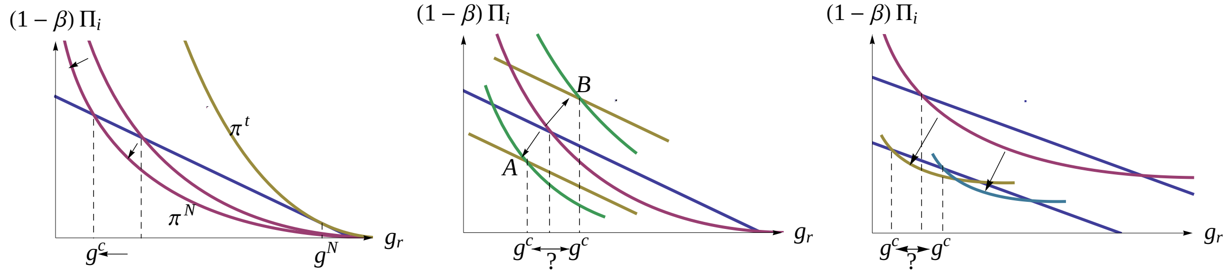
- $\frac{\partial Q^c}{\partial \beta} = 0$: *The discount factor has no effect on supply.*

Before finding the comparative statics on the level of violence, I assume that DTOs use optimal punishment, and therefore they achieve the best collusive equilibrium they can attain. The equation that determines the level of conflict is then the following IC:

$$\pi^c = (1 - \beta)\pi^t \tag{31}$$

The impact of policies can be found by determining how the level of violence has to adjust in order for this constraint to hold. The result is that the impact on the level of violence of exogenous changes depends entirely on how they affect the relative profits of collusion and betrayal. I will now discuss the results in an intuitive way, by using graphs of the profits. The formal proofs of these results are in appendix A.

Let us first consider an increase in β (shown in figure 3a). The profit from collusion remains the same, whereas the profit from betraying decreases: the DTO weighs the high one-period profit from betrayal less heavily. Since betraying becomes less appealing, g^c and G^N can decrease while still ensuring that treason is not profitable, which allows a lower level of violence.



(a) Increase in the discount factor.

(b) Increase in enforcement.

(c) Increase in the number of DTOs

Figure 3: Effect of different policies on individual expenditure in the conflict. Increasing the discount factor (3a) moves the curve of total profits when betraying (Π^t) down from the curve of the one-period profit from betraying (π^t), causing a decrease in the individual expenditure in the conflict. An increase in enforcement (3b) moves both curves either up or down, to point A or B in the figure, and profits from betraying move further away than profits from colluding. An increase in the number of cartels (3c) moves both curves down, and the net effect cannot be determined graphically.

An increase in enforcement causes either an increase or a decrease in profits in both situations (collusion and betrayal). Thus, it is not clear whether the incentives to betray increase or decrease (see figure 3b). I will now show that this is determined by the elasticity of demand. From the IC (31), the question is whether enforcement has a larger impact on π^c or $(1-\beta)\pi^t$, two quantities that were initially equal. This is equivalent to asking whether enforcement causes a larger percentage increase in profits from colluding or from betraying, since enforcement has no effect on $(1-\beta)$[27]. Thus, the quantities I want to compare are $\frac{\partial \pi^c}{\partial e}/\pi^c$ and $\frac{\partial \pi^t}{\partial e}/\pi^t$. The derivatives can be found from equation (25), which can be readily differentiated, and from applying the envelope theorem on (26):

$$\frac{\partial \pi^c}{\partial e} = \frac{\partial \pi^A}{\partial e} R^c \qquad \frac{\partial \pi^t}{\partial e} = \frac{\partial \pi^A}{\partial e} R^t \tag{32}$$

The envelope theorem thus implies that the change in profit is equal to the change in the profit they obtain from their productive behavior (i.e., ignoring expenditure in the conflict). Percent changes in productive profit are equal, since every DTO's productive profit is a constant fraction of the aggregate productive profit (π^A). Therefore, the impact of such changes on total profit depends on how effective colluding and betraying are in terms of how much DTOs spend in the conflict in order to obtain some share of productive profit. In order to betray, a DTO must increase its expenditure in the conflict by a significant amount, but since the marginal benefit of g is decreasing ($\frac{\partial^2 R}{\partial g_i^2} < 0$), the percent increase in routes is not as large as the percent increase in expenditure. This means that expenditure in the conflict is a larger share of the fraction of the productive profit it obtains than under collusion. Equivalently, final profit is a smaller share of productive profit, so a fixed percent change in productive profit causes a larger percent change in final profit when betraying than when colluding. This conclusion is proved in appendix A.

The question is thus whether enforcement increases or decreases productive profit π^A, just as in the analysis of enforcement at the SGNE (section 3.3). This depends on the elasticity of supply: there is a threshold $\hat{\epsilon}_c$ (given by equation (19)) such that enforcement increases productive profit, and violence, if $\epsilon_c > \hat{\epsilon}_c$. The opposite happens if $\epsilon_c < \hat{\epsilon}_c$.

The comparative statics as the number of DTOs changes is more complicated. An increase in the number of DTOs decreases both profits of colluding and betraying, since the routes must be shared among a larger number or DTOs. However, it is not clear which decrease in profits is larger (see figure 3c), so it is not clear whether individual expenditure increases or decreases. However, it can be shown that aggregate expenditure in the conflict, and therefore violence, increases. It is not a particularly illuminating process, so it is left to appendix A.

The following proposition summarizes the effect of different policies on violence:

Proposition 9. *If peaceful collusive equilibrium cannot be sustained (i.e. $\beta < \frac{n-1}{n}$), the comparative statics on the level of violence is as follows:*

[27] Equivalently, the question is whether $\frac{\partial \pi^c}{\partial e} \gtrless (1-\beta)\frac{\partial \pi^t}{\partial e}$. By using (31) in order to substitute for $(1-\beta)$ in terms of the profits, this becomes $\frac{\partial \pi^c}{\partial e}/\pi^c \gtrless \frac{\partial \pi^t}{\partial e}/\pi^t$.

- *If* $\epsilon_c < \hat{\epsilon}_c$*, then* $\dfrac{\partial G^c}{\partial e} < 0$*: If demand is sufficiently elastic, enforcement reduces violence.*
- *If* $\epsilon_c > \hat{\epsilon}_c$*, then* $\dfrac{\partial G^c}{\partial e} > 0$*: If demand is sufficiently inelastic, enforcement increases violence.*
- $\dfrac{\partial G^c}{\partial n} > 0$*: An increase in the number of DTOs increases the level of violence.*
- $\dfrac{\partial G^c}{\partial \beta} < 0$*: More forward-looking DTOs decreases the level of violence.*

The first three statements in proposition 9 (about $\frac{\partial G^c}{\partial e}$ and $\frac{\partial G^c}{\partial n}$) are not new: some very similar results were found for a SGNE. The intuition behind them, however, is very different, since DTOs' expenditure in the conflict is determined by a very different condition.

I already showed that the conflict over routes is equivalent to a conflict over the aggregate productive profit. Thus, what matters is whether enforcement increases or decreases productive profit. The crucial element that connects productive profit and violence is that changes in productive profit have a greater impact on betrayers than on colluders.

I argued in section 3.3 that the effect of enforcement on violence should focus on what happens with the marginal benefit of routes when enforcement increases. The reason for this is that violence was determined by a first order condition that equals marginal costs and benefits of expenditure in the conflict. Since the marginal benefit is proportional to the size of the prize being fought over, the question is again whether productive profit increases or decreases. The conclusion is that, in both cases, the effect enforcement has on violence is the same effect it has on productive profit, although the mechanism that translates higher productive profit into violence is very different.

In the SGNE, if violence remained equal after an increase in the number of DTOs, all DTOs would reduce their expenditure in the conflict, thus increasing the marginal productivity of expenditure. This means that DTOs increase their expenditure to stay at the optimum. The mechanism under collusion is very different. A greater number of DTOs means that the potential prize for a betrayer becomes greater: it can attempt to take away all other DTOs' routes. DTOs must therefore spend more in the conflict in order to deter potential betrayers, which leads to an increase in violence.

The result regarding the discount factor is new, as it played no role in a SGNE. It is perhaps the simplest result to grasp intuitively: more forward-looking DTOs are more fearful of punishment. This makes it easier to dissuade them with the threat of punishment, allowing DTOs to reduce deterrent expenditure in the conflict and decreasing violence, while still maintaining a collusive equilibrium.

3.4.3 Comparison with traditional collusion models

The model I just described shares many elements with collusion models from industrial organization (IO). In both cases, higher discount factors make it easier for collusion to hold. A lower number of players is also a facilitator of collusion in IO, as antitrust authorities are well aware: they check permanently bellwethers of collusion, the first of which are the number of firms operating in an industry and more advanced indicators of concentration such as the Herfindahl index.

A collusive agreement requires a high level of information in both cases. Players must hold two important pieces of information. First, they must know rivals' characteristics in order to determine how far down production (in IO) or expenditure in the conflict (in drug markets) can go while still fulfilling the IC, and providing the right incentives for others to collude. Obtaining this information is equally complicated in both cases. Second, players must monitor whether rivals comply with the collusion treaty in order to punish them if they do not. Otherwise, punishment is not a credible threat that deters deviation. In IO, this is very complicated since there is no way to know the quantity produced by each firm. The main source of information is the price of the good being sold, which can hint that somebody broke the agreement, but it says nothing about which was the precise firm that decided to deviate from collusion. On the other hand, monitoring is much easier in drug markets, since it is clear which DTO decided to increase resources spent in the conflict in order to increase its share of the routes.

Just as in IO, the theoretical model with perfect information predicts that players will always be in a collusive equilibrium, and they will never deviate since the IC is fulfilled. Thus, the games never reach the punishment stage. This is not the case in reality: firms wage price wars, and DTOs wage war against each other, sometimes eliminating rival DTOs. But this will clearly be the case once players have imperfect information. Uncertainty in the optimal level of production or expenditure in the conflict leads to violations of the collusive treaty: underestimates make it too easy for rivals to deviate, whereas overestimates can be misinterpreted as deviation. The error in measuring others' actions can also lead to punishing them when they actually complied with the treaty. These issues lead to the field of the economics of information, which deals with the means of communication and signaling that players may use in order to stay at an efficient equilibrium. A final point that applies to both cases arises from the economics of information: communication and coordination is harder for a larger number of players, which is an additional reason why more fragmented markets make it harder for collusion to hold. All these issues point to information and communication between DTOs as a potential extension to this model, in the same way as many models in IO were enriched by considering the possibility of imperfect information.

An important difference lies in the punishment strategies. Nash reversion is feasible and beneficial for players in both cases, but as argued by Abreu (1983), this is far from optimal. Punishment strategies in IO must take into account that when some firm deviates all others know that somebody deviated but they do not know who. Even if they knew who betrayed, they have no mechanism to punish the single firm that deviated. Abreu thus looked for punishments within *strongly symmetric* strategies, in which the punishment is equal for all firms, regardless of having betrayed or not. During the punishment phase, all firms tolerate some losses for the prospect of returning to the collusive agreement after a number of periods. But this complication is not necessary in drug markets, since which DTO betrayed is public information. Furthermore, DTOs can join forces against the lone traitor in order to give it an optimal punishment. Thus, the problem of giving an optimal punishment to deviators is much less complicated in the context of illegal drug markets than in IO.

There is perhaps an even more fundamental difference between my model and collusion in IO: collusive agreements in other industries are *negative* for society, since they move away from the efficient equilibrium that would be attained under perfect competition. On the other hand, collusion between DTOs is *positive* for society, since it decreases violence without having any effect on the productive behavior of DTOs. The role of governments is thus reversed: Antitrust authorities' main goal is to prevent collusive agreements, by checking industries whose characteristics make them more prone to collusion, by punishing colluders, and by implementing leniency programs that allow them to break existing agreements. Governments in drug trafficking nations should instead abstain from following policies that make collusion harder (see section 4).

3.5 Elasticity threshold for particular functional forms

Threshold (19) determines whether violence increases or decreases both in the SGNE and under collusion. I will now analyze the form that the threshold takes for some particular functional forms. If q is a Cobb-Douglas or CES function, the correction to the threshold vanishes[28]. However, these functional forms are not the most appropriate for this particular industry. With either a Cobb-Douglas or a CES function, holding the amount of drugs x fixed and increasing R results in $q(x, R, e)$ increasing without bound. But this makes no sense since the amount of drugs that reach the consumer market cannot be larger than the amount of drugs bought from the producer market, regardless of the amount of routes: this would mean a survival rate greater than one.

[28] For example, for a Cobb-Douglas function whose productivity parameter depends on enforcement, $q(x, R, e) = A(e)x^{\alpha}R^{1-\alpha}$. This results in $\frac{\partial \log q}{\partial e} = \frac{\partial \log \frac{\partial q}{\partial X}}{\partial e} = \frac{A'(e)}{A(e)}$, so the correction vanishes.

I will now consider another production function which better fits this particular conflict, a standard contest-success function (CSF) that takes into account both parties' resources to determine which fraction of the prize is obtained by each one of them[29]:

$$w(r, e) = \frac{r}{r + \varphi e} \tag{33}$$

The survival rate is now in the interval $(0, 1)$. The production function is $q(x, R, e) = xw(x/R, e)$. In terms of w, the threshold can be written as[30]

$$\hat{\epsilon}_c = -1 - \frac{\left(w - r\frac{\partial w}{\partial r}\right)^2}{r^2 \frac{\partial w}{\partial e} \frac{\partial^2 w}{\partial r^2}} \left(\frac{\frac{\partial w}{\partial e}}{w} - \frac{\frac{\partial w}{\partial e} - r\frac{\partial^2 w}{\partial r \partial e}}{w - r\frac{\partial w}{\partial r}} \right) \tag{34}$$

The sign of the correction is again determined by the sign of the term in parentheses.

Calculating the derivatives of w and substituting leads to the following expression for the term in parentheses: $\frac{\varphi}{r+\varphi e} > 0$. Thus, with this production function the effect of the marginal productivity is *always* larger than the effect on productivity, and the correction means that demand must not be too inelastic for enforcement to increase violence.

A simple expression for the threshold can be found by noting that homogeneity of degree zero of w means that r is uniquely determined by the ratio of the prices $\gamma = \frac{p_p}{p_c}$. The first order condition (5) is thus $w - r\frac{\partial w}{\partial r} = \gamma$. This results in a quadratic equation in r, whose solution leads to $r = \frac{\sqrt{\gamma}}{1-\sqrt{\gamma}}\varphi e$. All derivatives of w can now be rewritten solely in terms of e and φ, and substituting them in the threshold yields

$$\hat{\epsilon}_c = -\left(1 + \frac{\sqrt{\gamma}}{2(1 - \sqrt{\gamma})}\right) \tag{35}$$

Figure 4 shows the effect of varying γ. Its effect is the fact that as the gap between consumer and producer prices widens, and γ goes down to zero, costs become a smaller share of revenues and the threshold gets closer to -1. For high γ, the threshold goes well below -1, but empirical evidence shows that each step in the production chain of drugs, from producers to consumers, implies a large increase in prices (Mejía and Rico, 2010), and the results for high values of γ are therefore not very relevant.

Although the threshold does not depend directly on the level of enforcement, it depends indirectly through γ: as enforcement reduces supply, the price in the consumer market increases,

[29] The functional form that I present depends on the ratio of the resources committed by each party, and it is the most commonly used function in the literature of the economic theory of conflicts. Hirshleifer (1989) analyzes the implications of this type of function and some alternatives.

[30] The first two derivatives of r are $\frac{\partial r}{\partial x} = -\frac{R}{x^2}$ and $\frac{\partial^2 r}{\partial x^2} = \frac{2R}{x^3}$. All individual derivatives of q can be found from these two expressions and the chain rule: $\frac{\partial q}{\partial X} = w - r\frac{\partial w}{\partial r}$, $\frac{\partial q}{\partial e} = x\frac{\partial w}{\partial e}$, $\frac{\partial^2 q}{\partial X^2} = \frac{r^2}{x}\frac{\partial^2 w}{\partial r^2}$, $\frac{\partial^2 q}{\partial e \partial X} = \frac{\partial w}{\partial e} - r\frac{\partial^2 w}{\partial r \partial e}$. These can now be substituted in the expression for the threshold.

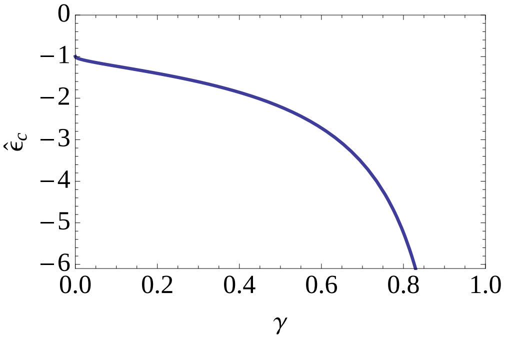


Figure 4: Variation of the threshold as a function of γ

causing a decrease in γ. Thus, as governments increase enforcement, they make it harder for enforcement to induce an increase in violence.

It may seem that the previous results depend on the particular functional form used for $w(r,e)$. However, I show in appendix B that the results are very similar for other functional forms fulfilling the condition that w cannot be greater than one: If the fraction of drugs reaching the consumer market is restricted to $(0,1)$, a few reasonable assumptions lead to the conclusion that the threshold for elasticity is lower than -1, and it is easier for enforcement to increase violence than predicted by previous works.

4 Enforcement, attacks on leaders, and fragmentation

The analysis that I have developed so far shows that there is a misconception regarding the traditional methods that policymakers have promoted in order to curb supply: they do not achieve their aim, while causing important increases in violence in trafficking countries. High-profile operations whose aim is to capture or kill bosses or to demobilize previously existing DTOs involve large political gains for governments, both because of popular support and because the international community, led by the U.S. government, has always promoted them. But like any policy that focuses on DTOs' military operations, they have no consequence on the quantity of drugs supplied to consumer regions, since aggregate productive behavior is independent of the war being fought over the control of routes, as shown in section 3.2.

These policies are not only inefficient; they are precisely the type of government action that may increase violence. Both the demobilization of previous organizations and successful attacks on cartel leaders create voids of power that are usually filled by more than one single group, thereby increasing the fragmentation of DTOs. Additionally, operations against bosses instill a feeling of restlessness and impatience in cartel leaders, who come to believe that their tenure is about to end. DTOs' strategies will then be focused on short-term operations that may bring temporally large profits, without much concern for future operations, and this is precisely the type of strategic planning that leads to higher violence. If the conditions are such that DTOs were initially able to form a tacit treaty in which each one controls some routes without any violence between them, fragmentation and impatience can trigger the breakdown of the peaceful equilibrium, after which drug trafficking becomes violent. And even if DTOs could not form a peaceful treaty at the beginning, fragmentation and impatience increase the level of violence under which they operate.

The previously mentioned policies may bring no benefits, but a large repertoire of alternative policies is available to governments. If their main purpose is to decrease the amount of drugs reaching consumer regions, as has been the case with traditional U.S.-led war on drugs, policies that reduce the productivity of DTOs, i.e., enforcement, fulfill their aim: patrolling routes and borders in order to increase the rate of seizures, attacking key lieutenants that coordinate the shipment of drugs through routes, and seizing assets used to transport drugs, such as boats, submersibles and airplanes. The supply reduction achieved by enforcement, however, may have an adverse effect. Drugs are addictive, which means that demand tends to be inelastic. Thus, supply reductions can cause large increases in the price of drugs, which increase potential profits by DTOs and encourage more violent operations. This results in a tradeoff for authorities willing to decrease supply: if they are successful, they will increase violence in trafficking regions. Previous works based on an incomplete analysis of DTOs' profits had shown that this adverse effect would take place if the elasticity of demand is higher than -1. I looked further into this number, by taking into account DTOs' costs, with a pessimistic conclusion: demand can even be elastic, while still allowing DTOs' profits and violence to increase as enforcement increases.

Some governments justify attacks on cartel bosses by arguing that such attacks disrupt the operations of DTOs, but it is not clear that attacking cartel bosses reduces the productivity of DTOs: past experience has shown that bosses can be readily replaced by former lieutenants with a seamless transition in the productive operations of DTOs. What matters is whether such attacks affect DTOs' productivity as drug traffickers, i.e., how efficient they are in taking the drugs they buy at producer nations to consumer nations through the land they control. Although drug leaders are a vital part of DTOs, they usually play a larger role in the conflict over routes, since they act as warlords. The advice of this paper is that authorities should focus on raids that have the largest impact on the productive operations of DTOs.

This discussion fits very closely the Colombian and Mexican cases I mentioned in the introduction. The government-led demobilization of the AUC in Colombia was followed by the emergence of a number of criminal bands, in a state of large fragmentation of DTOs. The areas where these bands operate have been among the most violent regions in Colombia in recent years. Mexican President Calderón's war against drugs is perhaps an even clearer example. Before his war, a few DTOs operated in Mexico in a state of peace. Mexican homicide rates were among the lowest in Latin America, despite the fact that most of the cocaine that went from Andean nations to the U.S. was shipped through Mexico. After Calderón started attacking drug leaders frontally, some DTOs broke into smaller pieces, and new DTOs were able to grab some portion of the illegal drug business. Most importantly, the level of violence doubled between 2006 and 2010. Both cases agree with an initial situation in which DTOs were patient enough and their number was low enough that they could collude in peace, but the government's actions induced the breakdown of the peaceful equilibrium and led to a new state of war.

This paper analyzes some key aspects of the interaction between DTOs, but in order to do so, it must inevitably leave aside some other important elements. Otherwise, the model would be too complex to solve and understand. I do not pretend this model to be interpreted literally as a picture of the real world, but this does not undermine the value of the advice it provides: it helps to understand some important mechanisms that are at work when governments plan their policy. I will now give one important example that illustrates this point. Probably the most important aspect that I do not consider are the dire consequences on politics and the rule of law that arise from a small number of DTOs holding a large share of the illegal drug trade in a trafficking region. This could be an important reason for governments to attack DTOs leaders, as Felipe Calderón did. Mexicans were well aware of the positive impact that the war on drugs would have by reducing the power held by cartel leaders. However, they had little idea that this war would bring the huge toll in deaths it caused. This paper can serve as a valuable piece of information to governments under similar circumstances, since it may compel them to adjust their policies once they are warned of the large negative effects they may have.

5 Conclusions

In this paper I extend the analysis of DTOs as single-period profit maximizers to a repeated-interaction approach. DTOs are modeled as firms that buy drugs at a producer region and attempt to take them to consumers through a trafficking region. In the process, they engage in two conflicts: they fight against other DTOs over who controls routes in the trafficking region, and they engage government forces who try to seize drugs on their way to consumers. If DTOs have perfect information, they will never be at the stage-game Nash equilibrium (SGNE) that previous works analyze. Instead, they collude by decreasing the amount of resources spent in the conflict against other DTOs, which results in less bloodshed than predicted by the SGNE. A peaceful equilibrium without any violence between DTOs can be sustained if there are only a few powerful cartels that are interested in maximizing the present value of their profits with a high enough discount factor.

DTOs' productive behavior (the amount of drugs bought from upstream markets and the amount of drugs sold to consumer markets) remains unchanged if governments attack cartel leaders or if DTOs are more fragmented; this is a consequence of the fact that productive behavior is independent of the conflict over routes. Thus, some traditional policies fostered by the U.S.-led war on drugs do not accomplish their purpose of curbing supply. As an unintended consequence, such policies increase violence between DTOs: they harm trafficking regions while attaining no positive effect on consumer regions. Governments do have the means to reduce supply: enforcement activities, focused on reducing the productivity of DTOs, decrease the amount of drugs reaching final markets. However, enforcement is not totally beneficial, as it increases drug prices. Hence, it increases DTOs' profits if demand is not too elastic, after which DTOs fight for higher stakes in trafficking regions with increased levels of violence. Previous analyses that only took into account DTOs' revenues suggested that this happens if demand is inelastic. By also taking costs into account, I present an improved criterion with the implication that enforcement may increase violence even if demand is elastic. Hence, governments willing to decrease supply face a tradeoff, since they may do so through enforcement, but this usually comes at the cost of increasing violence in trafficking nations.

References

Abreu, Dilip, "Repeated games with discounting: A general theory and an application to oligopoly.," *Ph.D. thesis, Princeton University*, October 1983.

_ , "Extremal Equilibria of Oligopolistic Supergames," *Journal of Economic Theory*, 1986, *39*, 191–225.

Baccara, Mariagiovanna and Heski Bar-Isaac, "How to Organize Crime," *Review of Economic Studies*, 2008, *75* (4), 1039–1067.

Bardey, David, Daniel Mejía, and Andrés Zambrano, "The endogeneous dynamics of crime structure: Heracles' lessons on how to fight the Hydra," *MIMEO, Universidad de los Andes*, 2013.

Becker, Gary S., Kevin M. Murphy, and Michael Grossman, "The Market for Illegal Goods: The Case of Drugs," *Journal of Political Economy*, 2006, *114* (1).

Bogliacino, Francesco and Alberto J. Naranjo, "Coca Leaves Production and Eradication: A General Equilibrium Analysis," *Economics Bulletin*, 2012, *32* (1), 382–397.

Burrus, Robert T., "Do Efforts to Reduce the Supply of Illicit Drugs Increase Turf War Violence? A Theoretical Analysis," *Journal of Economics and Finance*, 1999, *23* (3), 226–234.

Camacho, Álvaro, "Paranarcos y narcoparas: trayectorias delicuenciales y políticas," in Álvaro Camacho, ed., *A la sombra de la guerra. Ilegalidad y nuevos órdenes regionales en Colombia.*, Ediciones Uniandes - CESO, 2009.

_ , "Narcotrafico: mutaciones y política," in Alejandro Gaviria and Daniel Mejía, eds., *Políticas antidroga en Colombia: éxitos, fracasos y extravíos*, Ediciones Uniandes, 2011.

Castillo, Juan Camilo, Daniel Mejía, and Pascual Restrepo, "Illegal drug markets and violence in Mexico: the causes beyond Calderón," *MIMEO, Universidad de los Andes*, 2013.

Caulkins, Jonathan and Peter Reuter, "Redefining the Goals of National Drug Policy: Recommendations from a Working Group," *American Journal of Public Health*, 1995, *85* (1058-1063).

Chumacero, Rómulo A., "Evo, Pablo, Tony, Diego and Sonny: General Equilibrium Analysis of the Market for Illegal Drugs," *World Bank Policy Research Working Paper No. 4565*, 2008.

Garfinkel, Michelle R. and Stergios Skaperdas, "Economics of Conflict: An Overview," in Todd Sandler and Keith Hartley, eds., *Handbook of Defense Economics - Defense in a Globalized World*, Vol. 2, Elsevier, 2007.

Grossman, Herschel I. and Daniel Mejía, "The war against drug producers," *Economics of Governance*, 2008, *9* (1), 5–23.

Guerrero, Eduardo, "La raíz de la violencia," *Nexos*, 2011.

Hirshleifer, Jack, "Conflict and rent-seeking success functions: Ratio vs. difference models of relative success," *Public Choice*, 1989, *63* (101-112).

Lee, Li Way, "Would Harassing Drug Users Work?," *Journal of Political Economy*, 1993, *101* (5), 939–959.

Mailath, George J. and Larry Samuelson, *Repeated Games and Reputation*, Oxford University Press, 2006.

Mas-Colell, Andreu, Michael D. Whinston, and Jerry R. Green, *Microeconomic Analysis*, Oxford University Press, 1995.

Mejía, Daniel and Daniel Rico, "La microeconomía de la producción y tráfico de cocaína en Colombia," *Documento CEDE 2010-19 - Universidad de los Andes*, 2010.

_ **and Pascual Restrepo**, "The War on Illegal Drug Production and Trafficking: An Economic Evaluation of Plan Colombia," *Documento CEDE 2008-19*, 2008.

_ **and** _ , "The war on illegal drugs in producer and consumer countries: A simple analyticial framework," *Documento CEDE 2011-02 - Universidad de los Andes*, 2011.

Motta, Massimo, *Competition Policy: Theory and Practice*, Cambridge University Press, 2004.

O'Neil, Shannon K., "The Real War in Mexico: How Democracy Can Defeat the Drug Cartels," *Foreign Affairs*, 2009, *88* (4).

Poret, Sylvaine, "Paradoxical effects of law enforcement policies: the case of the illicit drug market," *International Review of Law and Economics*, 2003, *22*, 465–493.

_ **and Cyril Téjédo**, "Law enforcement and concentration in illicit drug markets," *European Journal of Political Economy*, 2006, *22*, 99–114.

Reuter, Peter and Mark Kleiman, "Risks and Prices: An Economic Analysis of Drug Enforcement," *Crime and Punishment*, 1986, *7.*

Tirole, Jean, *The Theory of Industrial Organization*, The MIT Press, 1988.

Appendix A Comparative statics for a collusive equilibrium

The total differential of (31) is

$$\left[\frac{\partial \pi^c}{\partial g_r} - (1-\beta)\frac{\partial \pi^t}{\partial g_r}\right] dg_r = -\pi^t d\beta + \left[(1-\beta)\frac{\partial \pi^t}{\partial e} - \frac{\partial \pi^c}{\partial e}\right] de + \left[(1-\beta)\frac{\partial \pi^t}{\partial n} - \frac{\partial \pi^c}{\partial n}\right] dn \tag{36}$$

which results in the following derivatives:

$$\frac{\partial g^c}{\partial \beta} = -\frac{\pi^t}{\frac{\partial \pi^c}{\partial g_r} - (1-\beta)\frac{\partial \pi^t}{\partial g_r}} \tag{37}$$

$$\frac{\partial g^c}{\partial e} = \left[-\frac{\partial \pi^c}{\partial e} + (1-\beta)\frac{\partial \pi^t}{\partial e}\right]\left[\frac{\partial \pi^c}{\partial g_r} - (1-\beta)\frac{\partial \pi^t}{\partial g_r}\right]^{-1} \tag{38}$$

$$\frac{\partial g^c}{\partial n} = \left[-\frac{\partial \pi^c}{\partial n} + (1-\beta)\frac{\partial \pi^t}{\partial n}\right]\left[\frac{\partial \pi^c}{\partial g_r} - (1-\beta)\frac{\partial \pi^t}{\partial g_r}\right]^{-1} \tag{39}$$

The term in the denominator is positive: it is the difference between the derivatives of the profit from colluding and the profit from betraying. I am analyzing the collusive equilibrium, at which the profit from colluding is less negatively sloped (see figure 2). The profit from betrayal is positive, so $\frac{\partial g^c}{\partial \beta} < 0$. Multiplying it by n yields $\frac{\partial G^c}{\partial \beta} < 0$.

The sign of (38) is the sign of its numerator. The analysis from 3.4 concludes that this is equivalent to comparing $\frac{\partial \pi^c}{\partial e}/\pi^c$ and $\frac{\partial \pi^t}{\partial e}/\pi^t$, i.e.,

$$\frac{\frac{\partial \pi^A}{\partial e} R^c}{\pi^A R^c - g^c} \quad \text{and} \quad \frac{\frac{\partial \pi^A}{\partial e} R^t}{\pi^A R^t - g^t} \tag{40}$$

From proposition 2, $\frac{x^c}{x^t} = \frac{R^c}{R^t} = \frac{q^c}{q^t}$, which allows me to multiply both the numerator and denominator on the right side by $\frac{R^c}{R^t}$, so the two quantitates to compare are

$$\frac{\frac{\partial \pi^A}{\partial e} R^c}{\pi^A R^c - g^c} \quad \text{and} \quad \frac{\frac{\partial \pi^A}{\partial e} R^c}{\pi^A R^t - \frac{R^c}{R^t} g^t} \tag{41}$$

The amount of routes in collusion and betrayal are $R^c = R(g_r, (n-1)g_r)$ and $R^t = R(g^t, (n-1)g_r)$. Since $R^t > R^c$, and R has decreasing marginal productivity, $\frac{g^t}{g^c} > \frac{R^t}{R^c} \implies \frac{R^c}{R^t} g^t > g^c$. This means that the denominator on the left side is greater than the one on the right side. The absolute value of the expression on the right side is thus greater, and in conclusion, if $\frac{\partial \pi^A}{\partial e}$ is negative, $\frac{\partial \pi^c}{\partial e}/\pi^c > \frac{\partial \pi^t}{\partial e}/\pi^t$, $\frac{\partial g^c}{\partial e} < 0$, and multiplying by n yields $\frac{\partial G^c}{\partial e} < 0$. If, on the other hand, $\frac{\partial \pi^A}{\partial e}$ is positive, all signs change, so $\frac{\partial \pi^c}{\partial e}/\pi^c < \frac{\partial \pi^t}{\partial e}/\pi^t$, $\frac{\partial g^c}{\partial e} > 0$, and $\frac{\partial G^c}{\partial e} > 0$.

The sign of (39) is undetermined, and it depends not only on the functional forms used, but also on the values of the parameters e, n and β. Not everything is lost, however, since I am primarily interested in finding the comparative statics on total violence, G^c. Thus, the derivative of interest is $\frac{\partial G^c}{\partial n} = \frac{\partial n g^c}{\partial n} = g^c + n\frac{\partial g^c}{\partial n}$. I will now show that it is positive, or equivalently, that $\frac{n}{g^c}\frac{\partial g^c}{\partial n} > -1$.

From the envelope theorem, the derivatives of profits are $\frac{\partial \pi^c}{\partial n} = p_c \frac{\partial q(x^c, R^c, e)}{\partial R_i} \frac{\partial R(g_r, (n-1)g_r)}{\partial g_{-i}} g_r$, $\frac{\partial \pi^t}{\partial n} = p_c \frac{\partial q(x^t, R^t, e)}{\partial R_i} \frac{\partial R(g^t, (n-1)g_r)}{\partial g_{-i}} g_r$, $\frac{\partial \pi^c}{\partial g^r} = -1$, and $\frac{\partial \pi^t}{\partial g_r} = p_c \frac{\partial q(x^t, R^t, e)}{\partial R} \frac{\partial R(g^t, (n-1)g_r)}{\partial g_{-i}} (n-1)$. Note that $\frac{\partial q(x^c, R^c, e)}{\partial R_i} = \frac{\partial q(x^t, R^t, e)}{\partial R_i}$, since $\frac{x^c}{x^t} = \frac{R^c}{R^t}$ (proposition 2) and q is homogeneous of degree one. From the first order condition for betrayal that relates the marginal benefit and cost of investment in the conflict, $p_c \frac{\partial q(x^t, R^t, e)}{\partial R_i} = \left[\frac{\partial R(g^t, (n-1)g_r)}{\partial g_i}\right]^{-1}$. Finally, since R is homogenous of degree zero, Euler's homogeneous function theorem means that $\frac{\partial R(g^t, (n-1)g_r)}{\partial g_i} = -(n-1)\frac{g_r}{g^t} \frac{\partial R(g_r, (n-1)g_r)}{\partial g_{-i}}$. Substituting all these expressions in (39) yields

$$\frac{n}{g_c}\frac{\partial g^c}{\partial n} = -\frac{\frac{\partial R^c}{\partial g_{-i}} - (1-\beta)\frac{\partial R^t}{\partial g_{-i}}}{\frac{n-1}{n}\left[\frac{g_r}{g^t}\frac{\partial R^t}{\partial g_{-i}} - (1-\beta)\frac{\partial R^t}{\partial g_{-i}}\right]} \tag{42}$$

where $\frac{\partial R^c}{\partial g_{-i}} = \frac{\partial R(g_r,(n-1)g_r)}{\partial g_{-i}}$ and $\frac{\partial R^t}{\partial g_{-i}} = \frac{\partial R(g^t,(n-1)g_r)}{\partial g_{-i}}$. By using Euler's homogeneous function theorem once again, $\frac{\partial R(g^c,(n-1)g_r)}{\partial g_i} = -(n-1)\frac{\partial R(g_r,(n-1)g_r)}{\partial g_{-i}}$, and from the IC, $(1-\beta) = \frac{\pi^c}{\pi^t}$. After some manipulation, this allows me to rewrite the last expression in terms of ratios between quantities for collusion and betrayal. For instance, $\tilde{g} = \frac{g^c}{g^t}$ is the ratio of expenditure in collusion to expenditure when betraying. Using a similar notation, $\tilde{R} = \frac{R^c}{R^t}$, $\tilde{\pi} = \frac{\pi^c}{\pi^t}$ and $\tilde{R}_i = \frac{\frac{\partial R^c}{\partial g_i}}{\frac{\partial R^t}{\partial g_i}}$. I thus obtain the following expression:

$$\frac{n}{g_c}\frac{\partial g^c}{\partial n} = \frac{n}{n-1}\frac{\tilde{g}\tilde{R}_i - \tilde{\pi}}{\tilde{\pi} - \tilde{g}} \tag{43}$$

I now want to express $\tilde{R}_i$ and $\tilde{R}$ in terms of $\tilde{g}$. In order to do so, recall that since the conflict is symmetric, $R(1, n-1) = \frac{1}{n}$, which means that $R(1,y) = \frac{1}{y+1}$ for any value of y. Homogeneity of degree zero of R means that $R(g^t,(n-1)g_r) = R(1,(n-1)g_r/g^t) = \frac{1}{(n-1)g_r/g^t+1}$, and after some manipulation, $\tilde{R} = \frac{R^c}{R^t} = \frac{n-1}{n}\tilde{g} + \frac{1}{n}$. On the other hand, $\frac{\partial R(g_r,(n-1)g_r)}{\partial n} = -\frac{1}{n^2}$, and from the chain rule $\frac{\partial R(g_r,(n-1)g_r)}{\partial n} = g_r\frac{\partial R(g_r,(n-1)g_r)}{\partial g_{-1}} = -\frac{g_r}{n-1}\frac{\partial R(g_r,(n-1)g_r)}{\partial g_i} = -\frac{1}{n-1}\frac{\partial R(1,n-1)}{\partial g_i}$. In the last two steps I used Euler's homogeneous function theorem and the fact that the derivatives of R are homogeneous of degree minus one. From the last expressions, $\frac{\partial R(1,n-1)}{\partial g_i} = \frac{n-1}{n^2}$, which means that $\frac{\partial R(1,y)}{\partial g_i} = \frac{y}{(y+1)^2}$. I am now in a position to find $\frac{\partial R(g_r,(n-1)g_r)}{\partial g_i} = \frac{n-1}{g_r n^2}$ and $\frac{\partial R(g^t,(n-1)g_r)}{\partial g_i} = \frac{1}{g^t}\frac{\partial R(1,(n-1)g_r/g^t)}{\partial g_i} = \frac{1}{g^t}\frac{(n-1)g_r/g^t}{((n-1)g_r/g^t+1)^2}$. By dividing both derivatives, I finally find that $\tilde{R}_i = \frac{(n-1+1/\tilde{g})^2}{n^2}$.

Note that $g^t > g^c$ and $\pi^t > \pi^c$, since the betrayer increases its expenditure in the conflict in order to increase its profits. Thus, both $\tilde{g}$ and $\tilde{\pi}$ are less than one. However, increasing expenditure in the conflict does not increase the profit proportionally, which means that $\tilde{g} < \tilde{\pi}$. This means that $\tilde{R}_i - \frac{\tilde{R}}{\tilde{g}} = \frac{(1-\tilde{g})n+\tilde{g}+\frac{1}{\tilde{g}}-2}{n^2\tilde{g}} > 0$, and $\frac{\tilde{\pi}-1}{n\tilde{g}} < 0$. Both expressions imply that $\tilde{R}_i > \frac{\tilde{R}}{\tilde{g}} + \frac{\tilde{\pi}-1}{n\tilde{g}} = \frac{n-1}{n} + \frac{\tilde{\pi}}{n\tilde{g}}$. Some straightforward algebra, in which one must be careful to change the direction of the inequality when dividing by $\tilde{g} - \tilde{\pi}$, yields $\frac{n}{n-1}\frac{\tilde{g}\tilde{R}_i-\tilde{\pi}}{\tilde{\pi}-\tilde{g}} > -1$. Therefore, $\frac{n}{g_c}\frac{\partial g^c}{\partial n} > -1$, and $\frac{\partial G^c}{\partial n} > 0$.

Appendix B Elasticity threshold for more general functions

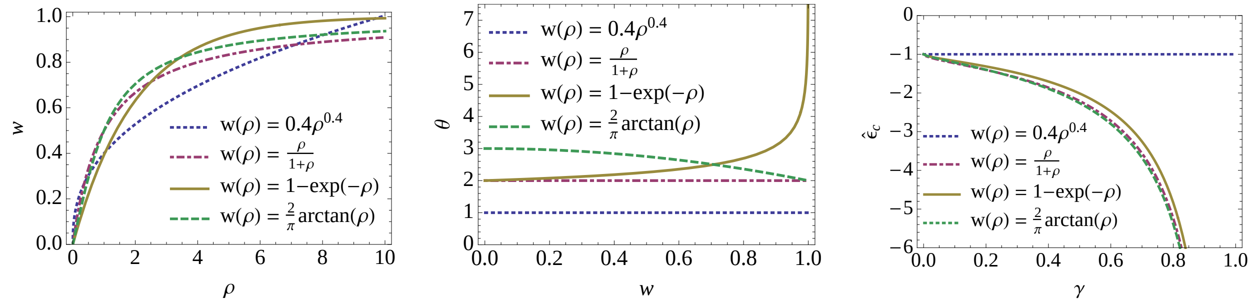
Suppose that w depends on the ratio of effective routes r to enforcement e. In order to allow for different efficiencies and increasing or decreasing returns to scale, I assume that w is a function of $\rho = \frac{r}{\varphi e^{\eta}}$: φ is a parameter that captures the relative efficiency of enforcement, and η is a parameter that captures whether the returns to scale of enforcement decrease faster than the returns to scale of effective routes. Thus, $w(r,e) = w(\rho)$. The conditions set on the derivatives of w in section 3.1 mean that $w' > 0$ and $w'' < 0$. This kind of function includes a variety of production technologies. For instance, if $w(\rho) = \rho^{1-\alpha}$ the production function is $q = e^{\eta(1-\alpha)}x^{\alpha}R^{1-\alpha}$, a Cobb-Douglas function, and the same CSF used in section 3.5 results if $w(\rho) = \frac{\rho}{1+\rho}$ with $\eta = 1$.

I will now show that such functions result in a correction that lowers the elasticity threshold $\hat{\epsilon}_c$. The term in parentheses in (34), which determines its sign, is now $\frac{rww''\rho_r\rho_e + rww'\rho_{re} - r(w')^2\rho_r\rho_e}{w(w - rw'\rho_r)}$. The denominator is positive, and by substituting the derivatives of ρ, its numerator is $-\rho ww'' - ww' + \rho(w')^2$, which is positive if

$$\theta = \frac{w''}{\frac{(w')^2}{w} - \frac{w'}{\rho}} > 1 \tag{44}$$

The numerator is clearly negative, and the numerator is also negative since the conditions on w imply that $w > \rho w'$. If (44) is fulfilled, the effect of enforcement on marginal productivity is greater than the effect on productivity, so the threshold is lower than -1. Condition (44) has the advantage that it is scale free: θ does not change by substituting $w(\rho)$ with $\hat{w}(\rho) = w(a\rho)$, where a is an arbitrary constant. It is also independent of η.

Setting $w_{CD} = \rho^{1-\alpha}$, a Cobb-Douglas technology, yields $\theta_{CD} = 1$. But as I argued in the main text, this is not a very reasonable form for w since it increases without bound. For it to be bounded above, given some value $w = w_{CD}$ and some value of $w' = w'_{CD}$, w'' should be less than for a Cobb-Douglas function ($w'' < w''_{CD}$) so that the function curves downward fast enough that it does not go past $w = 1$. This implies that $\theta > 1$. The relevance of θ being scale-free now becomes clear: the scale parameter a can be chosen so that $w = w_{CD}$ and $w' = w'_{CD}$, allowing comparison of θ and θ_{CD} only in terms of w'' and w''_{CD}.



(a) Different functional forms for $w(\rho)$

(b) Relation between derivatives (θ)

(c) Elasticity threshold

Figure 5: Comparison of different functional forms.

Figure 5 illustrates my argument graphically with three functions that fulfill the conditions for $w(\rho)$:[31] $w = \frac{\rho}{1+\rho}$, $w = 1 - \exp(-\frac{1}{2}\rho)$, and $w = \frac{2}{\pi}\arctan\rho$. I also show $w = 0.4\rho^{0.4}$ for comparison. The particular values of the parameters were chosen so that the functions are relatively similar,

[31] $w(0) = 0$, $w > 0$, $\lim_{\rho\to\infty} w(\rho) = 1$, $w' > 0$, and $w'' < 0$

although this does not change my conclusions. Figure 5a shows the general form of the functions. Figure 5b shows how θ behaves as a function of the value of w, and, in particular, that for all three functional forms $\theta > \theta_{CD}$. Finally, figure 5c shows the threshold that results for each functional form in terms of $\gamma = \frac{p_p}{p_c}$. Comparison with 4 shows that the conclusions from section 3.5 are not a peculiarity of the functional form that I chose for w.

Appendix C Varying prices in the producer market

In this section I relax the assumption that prices in the producer market are fixed. Since DTOs are price takers, their individual behavior does not change in any way, and their maximization problem is the same, both in the SGNE and with repeated games. The comparative statics, however, must now take into account that changes in policy will have an effect in the producer market, thus changing p_p. This effect is described by the elasticity of supply, which is now ϵ_p.

C.1 Aggregate productive behavior

From proposition 1, the number of DTOs has no effect on productive behavior, which means that it does not affect the amount of drugs bought from the producer region, and p_p. Thus, $\frac{\partial Q}{\partial n}$ stays the same. On the other hand, $\frac{\partial X}{\partial e}$ and $\frac{\partial X}{\partial e}$ do change. The analysis based on figure 1 is very similar, but it must now take into account that prices in producer markets are increasing in X, so marginal cost is increasing. The total differential analogous to (9) must now take this effect into account:

$$\left[\frac{dp_c}{dQ}\left(\frac{\partial q}{\partial X}\right)^2 + p_c\frac{\partial^2 q}{\partial X^2} - \frac{dp_p}{dX}\right]dX = -\left[\frac{dp_c}{dQ}\frac{\partial q}{\partial e}\frac{\partial q}{\partial X} + p_c\frac{\partial^2 q}{\partial X \partial e}\right]de \tag{45}$$

which results in the following expression, that replaces (10):

$$\frac{\partial X}{\partial e} = -\frac{\frac{\partial^2 q}{\partial X \partial e} + \frac{1}{Q\epsilon_c}\frac{\partial q}{\partial X}\frac{\partial q}{\partial e}}{\frac{1}{Q\epsilon_c}\left(\frac{\partial q}{\partial X}\right)^2 + \frac{\partial^2 q}{\partial X^2} - \frac{1}{X\epsilon_p}\frac{p_p}{p_c}} \tag{46}$$

The only change is a new term in the denominator, which does not change its sign, although it reduces its magnitude. From the chain rule, the new expression that replaces (12) is

$$\frac{\partial Q^e}{\partial e} = \frac{\frac{\partial^2 q}{\partial X^2}\frac{\partial q}{\partial e} - \frac{\partial q}{\partial X}\frac{\partial^2 q}{\partial X \partial e} - \frac{1}{X\epsilon_p}\frac{p_p}{p_c}\frac{\partial q}{\partial e}}{\frac{1}{Q\epsilon_c}\left(\frac{\partial q}{\partial X}\right)^2 + \frac{\partial^2 q}{\partial X^2} - \frac{1}{X\epsilon_p}\frac{p_p}{p_c}} \tag{47}$$

The sign of this expression does not change either. The comparative statics thus remains the same.

C.2 Threshold for the elasticity of demand

The effect of enforcement on violence depends on the effect it has on the aggregate productive profit. The new dependence of producer prices on quantities means that $\frac{\partial \pi^A}{\partial e} = \frac{\partial p_c}{\partial Q}\frac{\partial Q}{\partial e}Q + p_c\frac{\partial Q}{\partial e} - \frac{\partial p_p}{\partial X}\frac{\partial X}{\partial e}X - p_p\frac{\partial X}{\partial e}$. Rewriting $\frac{\partial p_c}{\partial Q}$ and $\frac{\partial p_p}{\partial X}$ in terms of elasticities leads to

$$\frac{\partial \pi^A}{\partial e} = p_c\left(1+\frac{1}{\epsilon_c}\right)\frac{\partial Q}{\partial e} - p_p\left(1+\frac{1}{\epsilon_p}\right)\frac{\partial X}{\partial e} \tag{48}$$

instead of (18). Substituting $\frac{\partial Q}{\partial e}$ and $\frac{\partial X}{\partial e}$ from (46) and (47) and isolating ϵ_c yields the following threshold for the elasticity of demand:

$$\hat{\epsilon}_c = -1 - \frac{\left(1+\frac{1}{\epsilon_p}\right)\left(\frac{\partial q}{\partial X}\right)^2}{\frac{\partial^2 q}{\partial X^2}\frac{\partial q}{\partial e} + \frac{1}{\epsilon_p}\frac{\partial q}{\partial X}\left(\frac{\partial^2 q}{\partial X \partial e} - \frac{1}{X}\frac{\partial q}{\partial e}\right)}\left(\frac{\frac{\partial q}{\partial e}}{Q} - \frac{\frac{\partial^2 q}{\partial X \partial e}}{\frac{\partial q}{\partial X}}\right) \tag{49}$$

Two new terms arise. First, the correction is smaller, since increasing marginal cost means that changes in X are smaller (the new term in the denominator)[32]. On the other hand, any change in X induces a larger change in costs, since p_p changes with X (see $\left(1+\frac{1}{\epsilon_p}\right)$ in the numerator). The sign of the correction is still determined by the sign of $\frac{\partial \log q}{\partial e} - \frac{\partial \log \frac{\partial q}{\partial X}}{\partial e}$.

[32] The sign of the correction could actually change if supply is very inelastic and $\frac{\partial^2 q}{\partial X \partial e} > \frac{1}{X}\frac{\partial q}{\partial e}$, but expanding this in terms of the derivatives of w shows that this would imply $\frac{\partial^2 w}{\partial e \partial r} > 0$.