

Optimal Timing of Use vs. Harm Reduction in an SA Model of Drug Epidemics

Jonathan P. Caulkins^{a,b},
Gernot Tragler^c, Dagmar Wallner^c

^a *Carnegie Mellon University, H. John Heinz III School of Public Policy and Management, 5000 Forbes Ave., Pittsburgh, PA 15213, U.S.A., email: caulkins@andrew.cmu.edu*

^b *Carnegie Mellon University in Qatar, PO Box 24866, Doha, Qatar, email: caulkins@andrew.cmu.edu*

^c *Department of Mathematical Methods in Economics, Vienna University of Technology, Argentinierstrasse 8/105/4, A-1040 Vienna, Austria. Email: (tragler@server.eos.tuwien.ac.at, dagmar.wallner@gmail.com, Corresponding author: +43-1-58801 11929 (phone), +43-1-58801 11999 (fax))*

Abstract

A debate in drug policy rankles between proponents of use reduction and harm reduction. We present a stylized two-state, one-control dynamic optimization model of this choice based on a social cost related definition of harm reduction, and parameterize it both for cocaine in the U.S. and for Australia's population of injection drug users. Static analysis of a binary choice between pure harm reduction and pure use reduction suggests that whether or not harm reduction is a good strategy can depend on various factors such as the particular drug, the country, the social cost structure, or the stage of the "epidemic". The optimal dynamic control version of the model involves boundary solutions with respect to the control variable with several switches in the optimal policy. The results have interesting interpretations for policy. Even for the U.S. parameterization, harm reduction turns out to have a potential role when drug use is either already pervasive or when use is so rare that there is no danger of explosive increases in initiation, but perhaps not when drug use is near a "tipping point". In contrast, in the parameterization for Australian IDU, where effective harm reduction tactics exist and budgetary cost for harm reduction measures are small, harm reduction appears preferable starting from any initial state. Furthermore, an interesting feature of our simple model is the occurrence of indifference curves, consisting of points where the decision maker is indifferent between two transients that will approach the same steady state in the long run. These transients result in the same social cost for the decision maker, but are characterized by quite different optimal policies.

Keywords: drug policy, harm reduction, optimal control, health, epidemics

1 Introduction

A central debate in drug policy concerns the relative merits of "use reduction" (meaning the control of drug use per se) versus "harm reduction" (focusing on reducing drug-related harms, e.g. reducing drug-related transmission of HIV).

There has been a similarly vigorous debate between proponents of "demand reduction" and "supply control". Previous work on optimal control of drug epidemics (Behrens et al., 1999, 2000, Tragler et al., 2001; Winkler et al., 2004) suggested a peaceful resolution to that dispute, namely that each had a crucial role to play, but with the relative emphasis varying over the course of the drug epidemic.

Inspired by those results, we explore here the possibility of a similar resolution to the harm reduction vs. use reduction discussion. Might the best policy not be either use reduction or harm reduction alone, but rather a blend with greater emphasis on use reduction at some stages of the epidemic and greater emphasis on harm reduction at other stages?

It is difficult if not impossible to investigate such a conjecture empirically. There simply are not enough countries with good measurement systems and with policies that shifted emphasis at different points in otherwise comparable epidemics to support a statistical analysis based on historical data. So instead we construct a stylized mathematical model of a drug epidemic, parameterize it for two different drugs/countries, and use it to simulate what might happen if policy were changed from use reduction to harm reduction or vice versa at different points in the country's trajectory of drug use.

There are many limitations to simulation models, most notably the inability to validate adequately either the model structure or the parameter values. Hence, we view this paper more as a thought experiment conducted subject to the discipline imposed by working within a formal mathematical model. We can speak definitively about what the best policy would be within the stylized world of our mathematical model and leave it to the reader to judge the extent to which any of the lessons carry over to the real world. We hope at a minimum that the question itself is interesting, that the results are provocative, and that the model provides a framework that helps readers subsequently discuss and debate the question of whether the relative merits of a use vs. harm reduction policy might vary over the course of a drug epidemic.

The next section refines the issue we seek to address and explains what we mean by drug epidemics. The third section translates these considerations into a formal simulation model expressed in differential equations. The fourth section describes simulation results when the decision maker gets to choose once and for all time whether the policy will be use reduction or harm reduction. The fifth section introduces an optimal dynamic control formulation that allows for blended strategies (varying policy along a continuous spectrum between pure use reduction and pure harm reduction) and the possibility that the optimal policy would switch emphases back and forth, rather than just in one direction.

2. Problem Definition

The term "harm reduction" is politically charged. While it is official policy in several countries (Australia, the Netherlands and the United Kingdom), it is denounced by U.S. drug policy makers as a deceitful ploy used by covert advocates of legalization. To complicate matters, different people use the term to mean different things. So the first order of business is to clarify what we mean by harm reduction.

MacCoun (1998) introduced a simple equation that concisely makes a critical distinction between the total harm caused by a drug (left side of Equation (1)) and the "harmfulness" of the drug, meaning the average harm per unit of drugs used.

$$\text{Total harm} = \text{Total use} * \text{average harm per unit of use} \quad (1)$$

In this paper we use the term "harm reduction" to mean reducing the harmfulness of drugs (rightmost term in Equation (1)).

We assume throughout that the ultimate objective of policy is to reduce total harm. This is akin to the standard objective of welfare maximization and, indeed, is identical if one ignores any possible benefits of drug use. Ignoring those benefits is inappropriate if the question under consideration is legalization, but since we are considering how to manage policy within a framework that makes the drug illegal, it is not unreasonable to ignore benefits people might derive from the inherently criminal activity of using drugs (Kleiman, 1992).

Total harm can be reduced in either of two ways, by reducing the quantity of drugs that is produced, distributed, and used (first term on the right hand side of

Equation (1)) or by reducing the harmfulness of the drug. The former is what we mean by “use reduction”; the latter is “harm reduction”.

What makes both use reduction and harm reduction controversial is the fear that driving down one term on the right hand side of Equation (1) might inadvertently drive up the other term.

That is, critics of use reduction argue that efforts to suppress use displace that use into more harmful forms. The classic example offered is that prohibition of syringe possession can lead injecting drug users to share and reuse their syringes, which increases the risk of infections and exacerbates the spread of blood-borne diseases, notably HIV/AIDS and Hepatitis C.

Critics of harm reduction argue that efforts to reduce the harmfulness of a drug might increase its use. This could happen because users and potential users respond to the objective changes in risk. Higher risks would deter use; lowering those risks reduces that deterrent. Or it could happen for symbolic reasons. E.g., the government distributing free needles or funding supervised injection facilities might be interpreted as an endorsement of drug use.

Either way, the basic tension underpinning our analysis is between reduced use and reduced harmfulness. We will embody that tension more specifically by building into the model a relationship between policies that reduce the harmfulness of current use and contemporaneous effects on initiation into drug use. That is, at least in our simple model, we do not worry about the possibility that harm reduction might slow exit from use; we focus only on potential adverse effects on initiation.

Is it reasonable to worry that harm reduction might increase drug use generally, and initiation in particular? People who believe human behavior is well described by rational actor models tend to say “of course”. Risks of adverse outcomes are seen as an additional cost of participating in an activity, albeit not one paid in dollars at the time of purchase. Nevertheless, lowering non-dollar costs might increase consumption in the same way that economists have shown that lowering the dollar costs of drugs increases drug use (Saffer and Chaloupka, 1999; Pacula et al., 2001; Williams, 2004; Grossman, 2005; Williams et al., 2006), even for outcomes related to dependent or heavy use (Bretteville-Jensen, 2006; Dave, 2006, 2008). Indeed, there is some empirical evidence that reducing enforcement risk does in fact lead to greater drug use (Farrelly et al., 1999; Desimone and Farrelly, 2003), some

evidence that adolescents' perceptions of the harmfulness of a drug helps explain trends in initiation (Bachman et al., 1998), and a long-standing belief that "search time costs" affect purchasing and use (Moore, 1977).

For people not trained in economics, the answer seems less clear. Yet, MacCoun (1998) discusses psychological and empirical evidence in various domains for the proposition that people tend to participate more frequently in a risky activity if that activity is made safer, a behavior that is sometimes called "risk compensation". Examples from the literature include the idea that drivers have responded to seat belts and other improvements in auto safety by driving faster and more aggressively, and that smokers compensate for low-tar cigarettes by smoking more cigarettes, inhaling more deeply, or blocking the filter vents.

The extreme version of risk compensation is "risk homeostasis" in which people increase their participation and/or reduce vigilance enough to fully offset the potential benefits of improved safety, in a sense adjusting behavior to make overall risk hit a subjectively acceptable "target risk" (Wilde, 1994). For example, the number of skydiving fatalities in the US has been stable (at around 30) for many years, even though skydiving equipment has become much safer (accident statistics from www.uspa.org).

MacCoun concludes that risk compensation happens, but not risk homeostasis in the sense that the proximate effects on use are generally less than proportionate to the change in risk. In particular, MacCoun (p.1203) concludes that there is "little evidence that behavioral responses produce net increases in harm [...]. Instead, most studies find that when programs reduce the probability of harm given unsafe conduct, any increases in the probability of that conduct are slight, reducing, but not eliminating the gains in safety [...]."

Consistent with MacCoun's conclusions, in our model risk compensation will act in the way that the direct or contemporaneous adverse effect of harm reduction on initiation is always less than the benefit resulting from reduced harm per unit of use. To foreshadow the model notation developed below, we will let the variable $0 \leq v \leq 1$ denote the reduction in harmfulness and $g(v)$ be a multiplier embodying the effect on initiation. It will always be the case that the percentage increase in initiation ($g(v) - 1$) is less than the percentage reduction in harm, i.e., $g(v) - 1 < v$ for all v .

That might seem to make the answer trivial (harm reduction always reduces total harm), but things can get more involved due to the second key concept underpinning our modeling, namely that drug use evolves according to a nonlinear dynamic process. The phrase “nonlinear dynamic process” though technically correct is opaque to most; the more common albeit sometimes misinterpreted term is a “drug epidemic”.

The concept of an “epidemic” merits explanation. Of course, there is no pathogen that spreads drugs use, the way that pathogens spread the flu, malaria or HIV. However, drug use is contagious in the same way fashions, laughter and even rumors can be (Noymer, 2001), and there is a long history of successfully applying epidemic models to describe trends in drug prevalence (Brill and Hirose, 1969; Hunt and Chambers, 1976; Rossi, 2001; Agar and Wilson, 2002; Caulkins, 2005).

Some worry that using the term “epidemic” will stigmatize users or implicitly justify draconian and inhumane interventions. However, the key idea is simply that drug use spreads through a diffusion process in which new users are primarily recruited by existing users (Golub and Johnson, 1996; Ferrence, 2001), as in classic models of new product adoption from marketing (e.g., Bass, 1969), rather than having drugs “pushed” on them by dealers (Coomber, 2006).

To the extent that initiation stems from social interaction between current users and current non-users, the value of preventing an initiation and, hence, the cost of inducing an additional initiation, can vary dramatically over the course of an epidemic (Winkler et al., 2004). That is, the “social multiplier” applied to a single exogenous intervention varies with the state of a drug epidemic (Caulkins et al., 1999, 2002). The principal justification for investigating the use vs. harm reduction choice within a dynamic model, rather than stopping with MacCoun’s Equation above, is the need to consider such feedback or multiplier effects.

The extreme version of a multiplier occurs when the drug “system” has multiple equilibria and the additional, marginal initiation can tip the system from a trajectory approaching one equilibrium to another trajectory approaching a different equilibrium. This notion of a “tipping point” is the third key concept underpinning our models. Discussion of tipping points have been in vogue since the publication of Malcolm Gladwell’s (2000) book by that name and Tom Schelling’s winning the 2005 Nobel Prize in economics for related work, but multiple equilibria separated by

tipping points have been a mainstay of models of drug markets and drug epidemics since well before that (Kleiman, 1988, 1993; Baveja et al., 1993; Caulkins, 1993).

The mathematical notion of a tipping point is more technical and a bit more precise than the informal notion Gladwell describes, but the essence is similar. In systems with nonlinear feedback, the long-run effects of a push to the system depends on what state the system was in at the time of that shove, and if the system happens to be very near the balance point between tipping one way or another, even a little shove can have very large effects on the outcome (“catastrophic” effects in the technical sense, not the alarmists sense, of the word).

Having introduced these three key concepts, the question we investigate can now be stated as follows. In a classic nonlinear drug epidemic model, can multiplier effects create situations – e.g., places near tipping points – where harm reduction can increase total harm even if the proximate effect of risk compensation on initiation is less than proportionate? To pursue this question further, we need a specific model, which we introduce next.

3. The Model

3.1 Model Overview

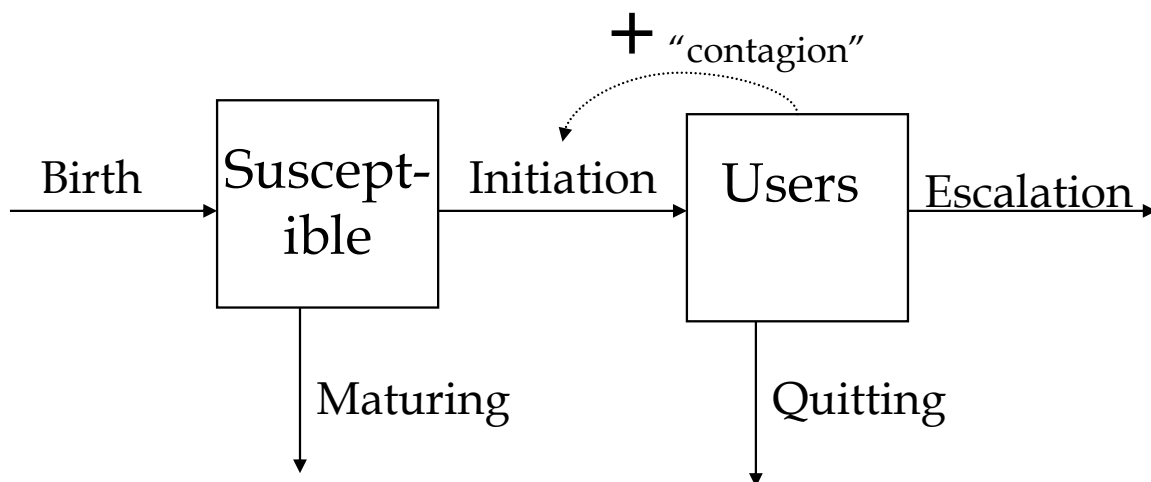
Perhaps the most commonly used model of drug use is the so-called LH model that differentiates between light and heavy users (Everingham & Rydell, 1994; Behrens et al., 1999, 2000; Caulkins et al., 2004). It is not ideal for our purposes for two reasons. First, when the LH model includes endogenous initiation, reputational feedback (the “Musto effect”) is a prominent driver of its dynamics, so it could be perceived to be biased against harm reduction. Second, we wish to contrast results for at least two different drugs/countries, but the LH model has only been parameterized for one country (the US) and some colleagues have argued that the Musto effect may be particular to cocaine in the US.

Instead we use what in many respects is a simpler two-state model of drug use, the “SA model” (Caulkins, 2004). The SA model is very much like the classic SIR model used in modeling infectious diseases. The main difference is that former (R = recovered) users are not modeled explicitly because they do not dilute interactions between never- and current users (S= susceptible and I = infected in SIR notation). That is, whereas a large pool of people who have been infected but since recovered

from some contagious disease (e.g., measles, rubella) slows the rate of new infection, the SA model does not imagine that a large number of former drug users discourages or otherwise interferes with initiation.

Figure 1 gives a schematic diagram of the SA model's two states and associated flows. One state, $A(t)$, tracks the number of drug users over time. The other, $S(t)$, tracks the number of people who are not now consuming drugs, but who are susceptible to initiating drug use.

Figure 1: Schematic Diagram of States in the SA Model



People enter the model in the susceptible state, $S(t)$, via a constant inflow rate, k , which can be understood as reaching an age when susceptibility to drug use starts. To be concrete, this could be young people turning age 12. After some ten or twenty years, those who had not yet tried drugs "mature out" of the S -pool, meaning they reach an age when they are no longer likely to start drug use. Age-structured models are possible, but considerably more complicated, so we do not have separate states for susceptibles who are age i for all i . Instead, we model this aging out process with a constant per capita outflow rate, δ , that is roughly equal to one over the usual dwell time in the pool S . E.g., if the duration of susceptibility is between ten and twenty years, thus we would expect δ to be between 0.05 and 0.1.

Exit from active use is likewise modeled via a constant per capita rate, μ , that includes death, ceasing use due to successful participation in a treatment program, or natural desistance.

The link between the two states is initiation. It is common in epidemiological models to assume that initiation is proportional to the number of contacts between users and non-users who are susceptible to initiation. In a random mixing model where infection-related deaths do not greatly alter the total population, the number of contact is itself just proportional to the product of the number of users and non-user, so it is usual to model initiation as:

$$\text{Initiation} = b A S, \quad (2)$$

for some positive constant b . We elaborate on that generic approach in two respects. First, Equation (2) applies when the probability that a random meeting between a user and non-user leads to an initiation is a fixed constant. That is a good model of biological infections, but not necessarily of social diffusion. Whether the product is a drug, a new fashion, or a new electronic gadget, the probability a non-user adopts a product after meeting a user can depend on how widely the product is used in the general population. So we modify Equation (2) by raising A to a positive constant, α , which may not necessarily be 1.

A classic justification for $\alpha > 1$ when it comes to new product adoption for consumer electronics are so-called “network externalities”, whereby the value to one person of using a product, such as a cell phone belonging to a certain network, increases when there are many other people on that same network (Shapiro and Varian, 1998). For illegal drugs, one can develop arguments why the initiation function should be concave ($\alpha < 1$) or convex ($\alpha > 1$) in the number of users.

For example, when a person is offered an illicit drug for the first time, he or she might be less likely to accept the offer if little is known about the drug and/or its use is highly deviant in the sense that only a small, and perhaps highly atypical subpopulation uses the drug. The person may be more likely to accept if many of his/her peers are already using it. If so, then we might expect a convex functional form with respect to A .

On the other hand, some adverse consequences of drug use do not manifest early on, when there are few users. They emerge later, when some people have used the drug for an extended period and when use is relatively widespread (Musto, 1987). In such cases, the virulence of the epidemic may decline as use spreads. Indeed, new initiations per current “light” user did decline for cocaine in the US as the cocaine epidemic progressed (Caulkins et al., 2004). An entirely separate argument that

points in the same direction is the idea of saturation. When use is relatively widespread, non-users may be offered the drug on multiple occasions. Suppose people who did not accept earlier offers are likely to reject future offers too. Then, as use grows, expansions in A lead to a less than proportionate increase in initiation. Either argument suggests a concave function $g(v)$ may also be appropriate for models of the use of certain drugs.

As we discuss below, best fitting parameterization suggest using a convex function for the U.S. cocaine epidemic ($\alpha > 1$) and a concave function ($\alpha < 1$) for Australia's population of injecting drug users (IDU).

The second elaboration on Equation (2) reflects the effects of harm reduction. Harm reduction will be modeled by a variable v , representing the proportion of harm to current users that is averted. For instance, if harm reduction interventions succeeded in reducing the harmfulness of drug use by one-third, then $v = 0.33$.

The central tension modeled in this paper is the possibility that harm reduction interventions might increase initiation, so we multiply initiation by a function $g(v)$. As far as we know, no one has ever empirically estimated such a function, but by definition, $g(0) = 1$, since when there is no harm reduction, initiation should remain at its baseline level. Also, in light of McCoun's (1998) argument that there may be risk compensation but not risk homeostasis, we require

$$1 < g(v) < 1+v \text{ for all } v > 0. \quad (3)$$

One functional form for $g(v)$ that meets these requirements can be derived by imagining that changes in the non-monetary cost of drug use experienced by the user (health risks, e.g.) affect initiation in ways that parallel the way that changes in the monetary cost of drug use do. This assumption is convenient because there is a growing empirical literature that deals with how responsive drug use is to changes in drug price (e.g. Grossman, 2004, Dave, 2006, 2008).

Before proceeding, it is important to distinguish the three types of drug-related costs: monetary costs, social costs, and non-monetary costs of drug use. The monetary cost of using a drug is just the purchase price the user pays to the drug supplier. The social costs are the total costs to society associated with the production, distribution, consumption, and control of the drugs; they are what the policy maker cares about. In the parameterization here, social cost estimates are derived from "Cost of Illness" (COI) studies (Single et al., 2003).

The personal, non-monetary costs are smaller than the social costs of drug use because some social costs are externalities from a user's point of view. Reducing them cannot increase initiation via risk compensation because, by definition, those externalities never enter into the users' decision about whether or not to try a drug.

The proportion of social costs that are internal indirectly places an upper bound (v_{\max}) on the control variable, v , in the following sense. Within the context of this model, reducing harms that are externalities from the user's perspective is a clear win with no risk of unintended consequences. This model cannot improve on that simple and unambiguous policy prescription, so we assume policy makers already reduce those external costs to the extent possible. Hence, the only harm reduction we consider within this model, is reductions in harmfulness to users. Again, that is not because reducing harms borne by non-users is unimportant; clearly it is. We do this simply because the policy recommendation to reduce harms to non-users is so obvious that we do not need a fancy model to support that recommendation.

As a proxy, we will assume that the health related costs documented in COI studies are borne by the users, and the other cost components are seen as externalities by the users. Of course some health-related costs are not borne by the user (e.g. emergency care provided pro bono), and some non-health costs are internal (e.g., reduced labor market productivity), so this assumption is great simplification. To the extent that the reader believes the former bias is larger (smaller) than the latter, the reader might be interested in sensitivity analysis exploring smaller (larger) values of the v_{\max} .

We said above that non-dollar costs borne by the user will be considered in parallel with dollar costs, but we will not assume they are weighted equally by the user. So the total cost of use as perceived by the user will be the monetary cost (price of the drug) plus a proportion ω of the drug-related social costs that are borne by the user. For example, consider an adverse health outcome that would occur fifteen years after initiation. If the social planner (decision-maker) used a 5% discount rate, but the user were more present-oriented and used a 10% discount rate (plausible, given Kirby and Petry, 2004), then in present-value terms the user would only weight that event half as heavily as the social planner would.

This parameter is important. If $\omega = 0$ then reducing harmfulness to users is like reducing harmfulness to non-users in the sense of creating no risk of increased

initiation. However, if $\omega = 1$, then the same harm reduction program would induce greater risk compensation than if ω were smaller. In the absence of empirical evidence, we take $\omega = 0.5$ as our baseline value, but fully acknowledge that different people may have different judgments about what value is most appropriate. Indeed, some people who disagree about the relative merits of use vs. harm reduction may disagree about the policy recommendation precisely because they have different judgments about how large ω is. If the model helps them shift their disagreement from one couched in terms of values or policy recommendations into one framed in terms of an objective parameter that at least in principle can be measured, that would be an example of the model's mathematics helping to clarify a debate even if the model cannot "answer" the question.

Letting c_m and c_s denote the monetary and social cost per unit of consumption, respectively, the forgoing implies that the full cost of use as recognized by the user when there is no harm reduction is $c_m + \omega v_{max} c_s$, whereas harm reduction v reduces that to $c_m + \omega (v_{max} - v) c_s$.

A common way economists model the effect of changes in price on changes in consumption behavior is through an elasticity, defined as the percentage change in consumption associated with a 1% increase in price. In complicated models of demand, the elasticity can be different over different price ranges, but when the price changes are modest, a simple constant price elasticity model may suffice. A constant elasticity model in this context would imply that

$$g(v) = \left(\frac{c_m + (v_{max} - v)\omega c_s}{c_m + \omega v_{max} c_s} \right)^\gamma. \quad (4)$$

Note: the exponent γ is the elasticity of participation with respect to the *total cost of using drugs*, as recognized by the user. Dave (2006, 2008) and others have empirically estimated the elasticity of participation with respect to the *price*, c_m , when c_m varies. It is a simple calculus exercise to determine how to convert one into the other. Table 1 below gives both the elasticity with respect to price, η , estimated based on the literature, as well as the corresponding implied value of γ .

Hence we use for initiation not Equation (2) above, but rather

$$Ini(A, S, v) = b A^\alpha S g(v) = b A^\alpha S \left(\frac{c_m + (v_{max} - v)\omega c_s}{c_m + \omega v_{max} c_s} \right)^\gamma. \quad (5)$$

For both the U.S. cocaine and Australian IDU parameterizations the function $g(v)$ is convex, albeit only modestly. Convexity is sensible since among the various programs, a smart policy maker would implement first programs that generate the biggest reduction in harm per unit impact on initiation.

If $g(v)$ were linear, the nature of the dynamics would lead to optimal policies that involve what are called “bang-bang controls”, meaning abrupt shifts from the smallest possible to the largest possible value of v . Our $g(v)$ is not linear, but with the parameters below, the convexity (curvature) is modest. This leads in the continuous control model (Section 5) to recommendations for rapid policy shifts and spending quite a bit of time at the boundary solutions of pure use reduction ($v = 0$) or pure harm reduction ($v = v_{\max}$). Because of that character, the simpler static model considered in the fourth section is more relevant than would otherwise be the case.

3.2 Formal Statement of Model

The SA model described above can be expressed concisely via the following system of two linked, nonlinear differential equations:

$$\begin{aligned}\dot{S} &= k - \delta S - bA^\alpha Sg(v) \\ \dot{A} &= bA^\alpha Sg(v) - \mu A\end{aligned}\tag{6}$$

The performance metric or outcome of interest is the total discounted drug-related social costs over some planning horizon (assumed to be infinite so no salvage value function needs to be specified). The control or policy variable $v(t)$ is the percentage reduction in the harmfulness of drug use. Social costs are the product of (1) the number of drug users, $A(t)$, (2) a baseline social cost per user per unit time when there is no harm reduction, which is normalized to 1 without loss of generality, and (3) one minus the proportion of harm that is averted via harm reduction policies, i.e., $1 - v(t)$. Hence, in symbols the objective function is:

$$Z = \int_0^{\infty} e^{-rt} (A(1-v)) dt\tag{7}$$

One could also penalize program spending on harm reduction by adding a $c(v)$ term to the objective function. However, harm reduction programs receive very modest levels of funding even in countries such as Australia and the Netherlands that make harm reduction the centerpiece of their national policies (Moore, 2005; Rigter, 2006). It is better to think of the control not as a program with a budget but rather as a

policy. For example, when a jurisdiction pursues the harm reduction policy of telling police not to arrest people for possessing a syringe, that policy costs essentially nothing more than the paper on which the new policy memos are printed. So we do not consider such a cost term $c(v)$ here.

3.3 Parameterization

Table 1 summarizes the parameter values derived in Caulkins et al. (forthcoming) for the SA model of the Australian IDU and US cocaine epidemics.

Table 1: Parameter Values

Parameter	Symbol	Australian IDU	US Cocaine
Inflow into S state	k	0.0526	1.3417
Exit from S state	δ	0.0952	0.0605
Coefficient in $Ini(t)$	b	0.5112	0.0090
Exponent in $Ini(t)$	α	0.8622	1.5604
Exit rate from A state	μ	0.1136	0.1661
Social cost of use	c_s	\$39,225/yr	\$223.56/gm
Proportion of social cost HR can avert	v_{max}	53%	17.408%
Proportion of health costs internalized	ω	0.5	0.5
Monetary cost of use	$\chi\mu$	\$13,537/yr	\$106.54/gm
Price Elasticity of participation	η	-0.21	-0.45
Cost elasticity of participation	γ	-0.371	-0.532
Annual discount rate	r	0.04	0.04

Structurally the most important parameter is the exponent, α , of the exponent in the initiation function, Equation (5). Since $\alpha > 1$ for US cocaine, initiation is a convex function of A for $A > 0$, and it is possible to have multiple stable equilibriums separated by a tipping point. In contrast, since $\alpha < 1$ for Australian IDU, there can be only one steady state with a positive amount of drug use, and it is stable. The absence

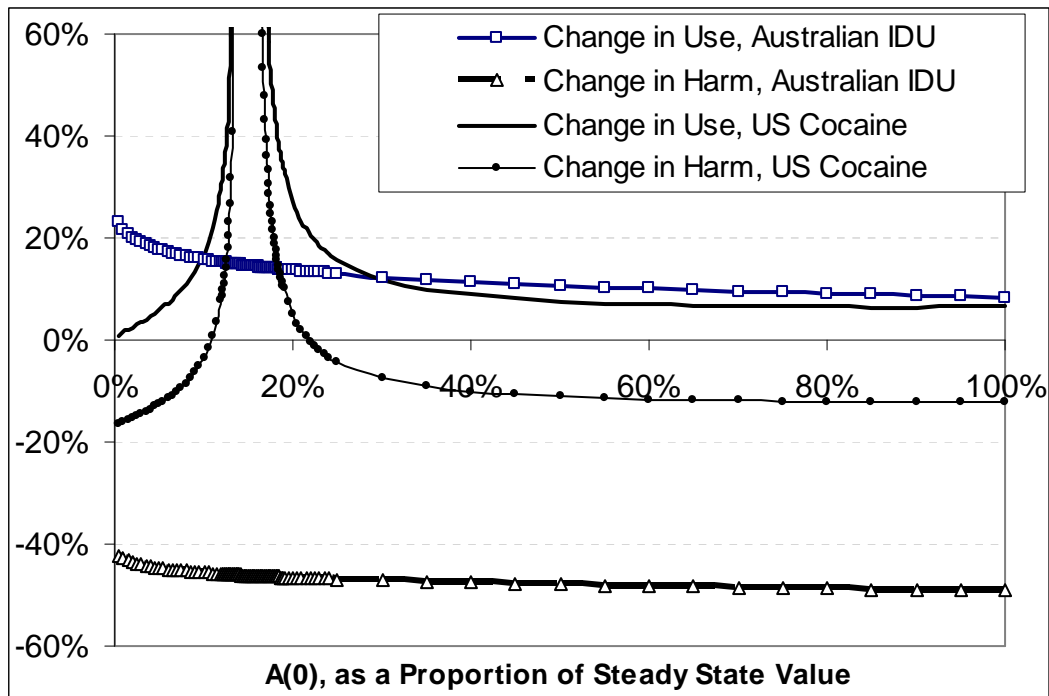
of a tipping point for Australian IDU implies that harm reduction has greater potential to trigger catastrophic increases in use for US cocaine.

The potential benefit of harm reduction is greater for Australian IDU because its upper bound, v_{max} , is much larger (0.53 vs. 0.174 for US cocaine). That reflects the fact that much of the harm associated with Australian IDU comes from outcomes for which effective harm reduction tactics exist (primarily preventing overdose and the spread of blood-borne infectious diseases), whereas more social costs associated with US cocaine pertain to crime, violence, and reduced labor productivity.

4.0 Analysis of One-Shot Policy Choices

With both the US cocaine and Australian IDU parameterizations, we computed the present value of all future use ($A(t)$) and harm ($A(t)(1 - v)$) for various initial numbers of users ($A(0)$), setting the initial number of susceptibles to be the corresponding steady state value ($S(0) = k - \mu A(0) / \delta$). Inasmuch as $A(t)$ tends to be increasing in the early stages of a drug epidemic, this essentially models different points in time at which harm reduction could begin. Figure 2 shows the results, with $A(0)$ expressed as a proportion of its positive steady state value when there is no harm reduction, so results for both models can be shown on the same graph. (Otherwise numbers of US cocaine users are much larger than are numbers of Australian IDUs.)

Figure 2: Comparison of Effects of Implementing Harm Reduction with Different Initial Numbers of Users for US Cocaine and Australian IDU Epidemics



For the Australian IDU parameterization, regardless of $A(0)$, implementing harm reduction increases the (present values of) numbers of users but reduces aggregate harm. Unless $A(0)$ is quite small, implementing harm reduction increases drug use by 8-20% and reduces harm by 44-49%.

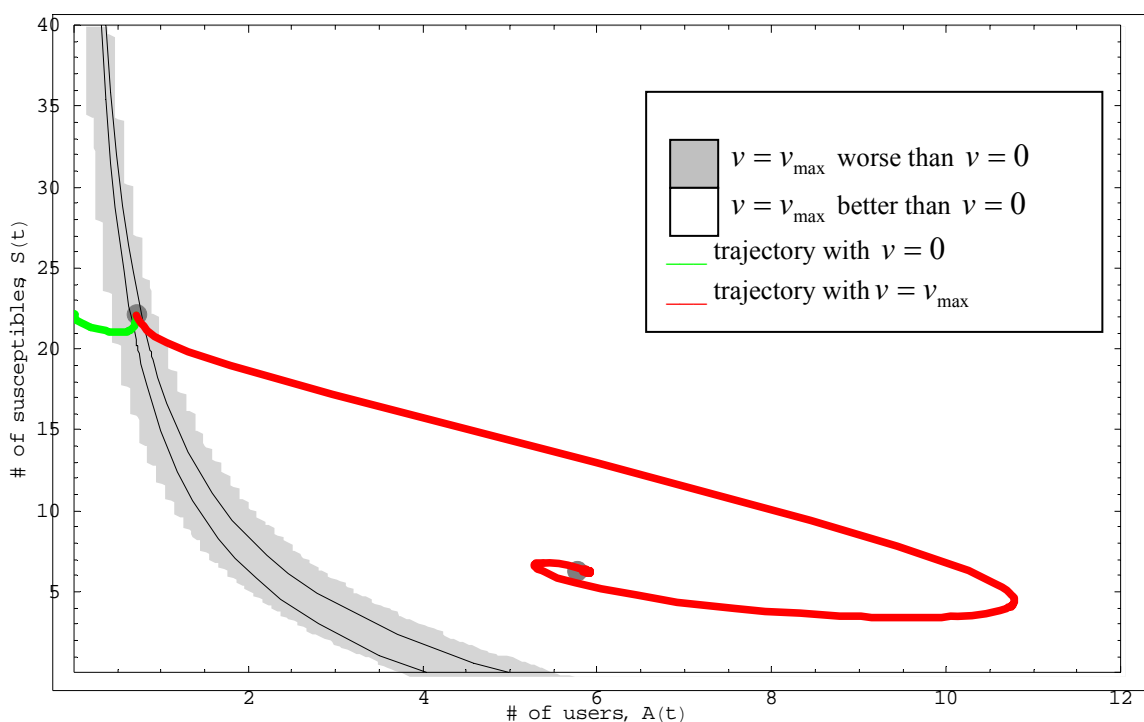
Results for US cocaine are qualitatively similar for $A(0)$ greater than about 30% of the steady state value, with harm reduction increasing use by 7-12% and reducing harm by 8-12%. Likewise, at the other extreme, for $A(0)$ less than about 11% of the steady state number of cocaine users, harm is reduced even though use goes up. However, for $A(0)$ between 11% and 22% of the positive steady state value (roughly 600,000 – 1,200,000 users), implementing harm reduction increases not only drug use but also total harm, dramatically so for certain $A(0)$. For $A(0)$ around 15% of the steady state value, total harm can be increased by more than a factor of 4.5.

The reason for these divergent results is that for the Australian parameters, there is only a single steady state that all trajectories approach, regardless of where they start. Implementing harm reduction shifts that steady state to the right (more users) and generally leads to greater use on the transient paths approaching the steady state, but the trajectories with and without harm reduction ($v = v_{\max}$ or $v = 0$) are not so dramatically different. (Figure not shown.)

In contrast, the US cocaine parameters generate a phase portrait (Figure 3) with both a low-level equilibrium and a high-level equilibrium (with $A = 5.5$ million

users) that are separated by a curve made up of so-called tipping points. Figure 3 shows where the tipping point curves are both with and without harm reduction. The right hand curve is the set of tipping points when there is no harm reduction ($v=0$). For initial conditions to its left, the system will converge to the low-use equilibrium, whereas initial conditions to the right of the curve lead to the high use steady state. The left hand curve shows the corresponding set of tipping points when harm reduction is implemented as aggressively as possible ($v = v_{max} = 17.4\%$).

Figure 3: Epidemic Trajectories in the S-A Plane for US Cocaine Parameterization with and without Harm Reduction, Starting at $A(0) = 717,691$ and $S(0) = 21,100,000$. Vertical curves show tipping points separating regions of convergence to low- vs. high-Levels of Use, both with (left hand curve) and without (right hand curve) harm reduction.



Implementing harm reduction shifts to the left the curve of tipping points (technically, the “separatrix”). Consider the implications of this for an initial point between the two curves, such as the point $A(0) = 717,691, S(0) = 21,100,000$. Without harm reduction, this point is on the left of the separatrix, so drug use will converge to

the low-level steady state. But if harm reduction is implemented, that point is on the right-hand side of the (now shifted) separatrix, so use will converge to a high-level equilibrium with 5.78 million users. Figure 3 illustrates this by contrasting the trajectories emanating from that point both with harm reduction (the curl down to the lower right) and without harm reduction (the short segment moving up and to the left toward the vertical axis). More generally, for any combination of A and S that lies between the two separatrices, harm reduction tips the model to the high-level equilibrium.

That harm reduction – even the relatively modest sort available for US cocaine where v_{max} is only 17.4% – could so terribly affect the trajectory of a drug epidemic might seem to be a sobering caution against using harm reduction. On the other hand, for this model and parameters, implementing harm reduction could adversely tip the epidemic only if the current state were within a quite small region in S - A space, namely the region between the tipping point curves. Furthermore, with those initial conditions, prevalence is relatively low and declining, so there might not be a lot of popular or political pressure to alter policy and implement harm reduction.

On the other hand, changing the parameter values can slide that critical region to the left or to the right. Since in practice those parameters will never be known with any precision, a decision maker would never know for sure whether or not a particular drug system were in that critical region or not. So even if the likelihood of adversely tipping the epidemic were low, it is not a possibility that can easily be ruled out when the system has multiple equilibria.

Furthermore, the region where harm reduction increases total harm (shaded gray in Figure 3) is broader than the small sliver where it tips the epidemic toward the high-level steady state. For $A(0)$ in these shoulder regions, implementing harm reduction does not tip the epidemic, but it does adversely affect the transient so severely that it increases the present value of total harm.

The fundamental insight from this discussion is that the benefit or cost of implementing harm reduction is “state-dependent” to use the technical term. That is, it depends on the state of the system, which in an SA model is defined by the number of users and the number who are susceptible to initiation.

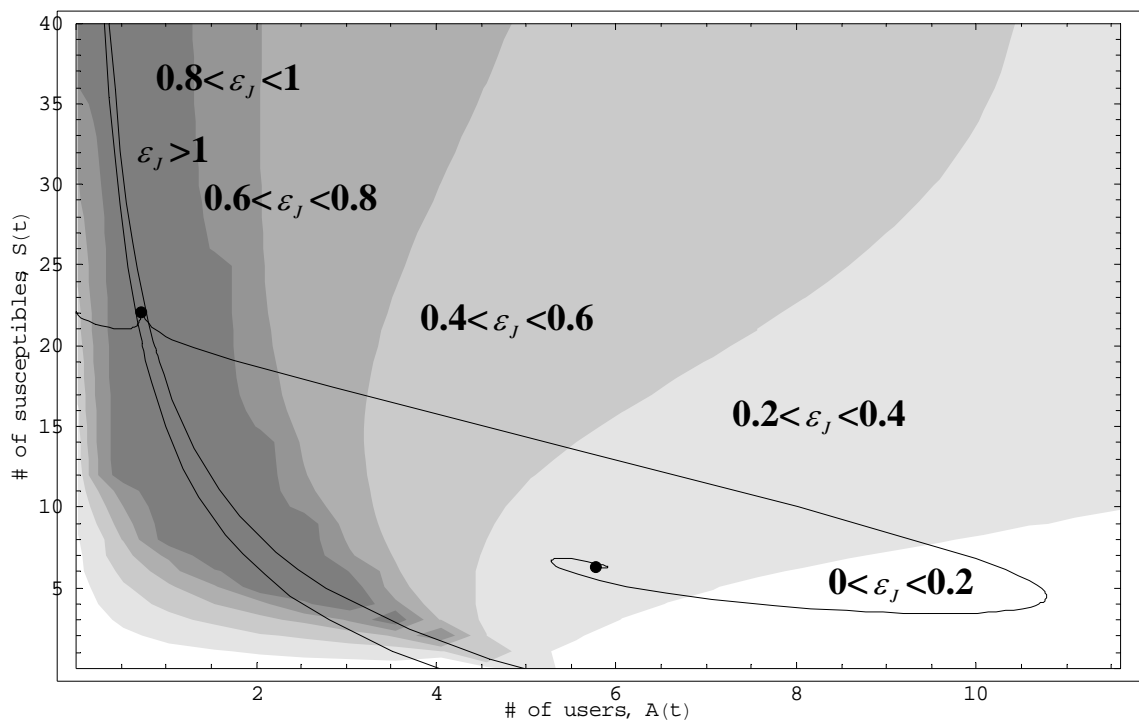
The underlying reason for this is that the consequences of having greater initiation is itself state dependent (Winkler et al., 2004). In this model, full-scale

harm reduction increases initiation by a factor of $g(v_{\max})$, which is 1.09 in the U.S. parameterization, implying that initiation is 9% higher than it otherwise would be from that point on.

Figure 3 only shows when a permanent $\sim 9\%$ increase in initiation increases drug use by more or less than $v_{\max} = 17.4\%$. So it answers the policy question, “If one has to decide today between choosing harm reduction now or never and with no chance to subsequently adjust policy, what should one do?”

Figure 4 provides a richer and more flexible look at the implications of pursuing a policy that might increase initiation, even temporarily. In particular, it explores the effect on the present value of all future drug use of increasing initiation by 1% for a single year, i.e., of increasing $g(v)$ from 1.0 to 1.01 for a single year and then returning to its base case value of 1.0. Figure 4 shows a contour plot of this elasticity (numbers multiplied by 1000), where the darker the color is, the worse is the effect of this temporary increase in initiation.

Figure 4: Elasticity ε_j of the present value of future drug use with respect to an increase in initiation by 1% for one year. The darkest gray region depicts initial conditions for which $\varepsilon_j > 1$, this is initial conditions for which the effect of the temporary increase in initiation is worst.



5.0 Analysis of An Optimal Dynamic Control Formulation

In the previous section we contrasted a stark policy choice of choosing forever between $v=0$ and $v=v_{\max}$, what might be called pure use reduction and pure harm reduction. Another and more general way to ask the policy question is, “For any given state of the system, how much harm reduction is optimal to use if one has complete flexibility to with regard to changing $v(t)$ over time?” Answering that question requires a more advanced type of mathematics, called optimal control theory. The mathematical formulation of this more general problem becomes

$$\begin{aligned} & \underset{0 \leq v \leq v_{\max}}{\text{Min}} \int_0^{\infty} e^{-\rho t} (A(1-v)) dt & (7) \\ \text{s.t.} \quad & \dot{S} = k - \delta S - bA^\alpha Sg(v) \\ & \dot{A} = bA^\alpha Sg(v) - \mu A \end{aligned}$$

To save space, we will not take the reader through the derivation and analysis, but Figure 5 presents the results from the optimal control model on the same type of S-A phase diagram as used above, Figure 5a for Australian IDU and 5b for US cocaine. As before, for any point on the diagram, the initial number of users $A(0)$ is specified by the horizontal axis coordinate and initial number of susceptibles $S(0)$ by the vertical axis. Pursuing the optimal policy will lead the epidemic to follow the curve passing through that point until it approaches a steady state. (There is only one steady state in Figure 5a for Australian IDU; in Figure 5b the curves approach either a low-use or a high-use steady state, as in Figure 3.) The colors along the path indicate how much harm vs. use reduction should be used at that point along the trajectory. In the regions colored light gray $v^* = 0$ (pure use reduction) is the optimal policy. Black indicates places where the optimal policy calls for $v^* = v_{\max}$ (pure harm reduction). In between, inner control values are optimal, the respective regions are colored in dark gray.

For Australia, for most of the initial conditions $v^* = v_{\max}$ is the optimal strategy, whereas for the U.S. case the state space is divided in a more complicated way. Figure 5 b) depicts the optimal policy for the U.S. cocaine epidemic. Colors are chosen in the same way as described above. Note, harm reduction is optimal in regions where the problem is still of modest size and there is less danger of explosions in initiation via the feedback-effect. Harm reduction also has significant merits when the

drug is already in pervasive use. Although the feedback effect curbing initiation is stronger in that region, the cost cutting effect in the objective function has a stronger overall effect. Thus, even a modest size harm reduction capability with $v_{\max} = 17\%$ would have merits for the present of the U.S. cocaine problem within the context of this simple model.

Figure 5 a) S-A-plot for optimal control strategies for Australia

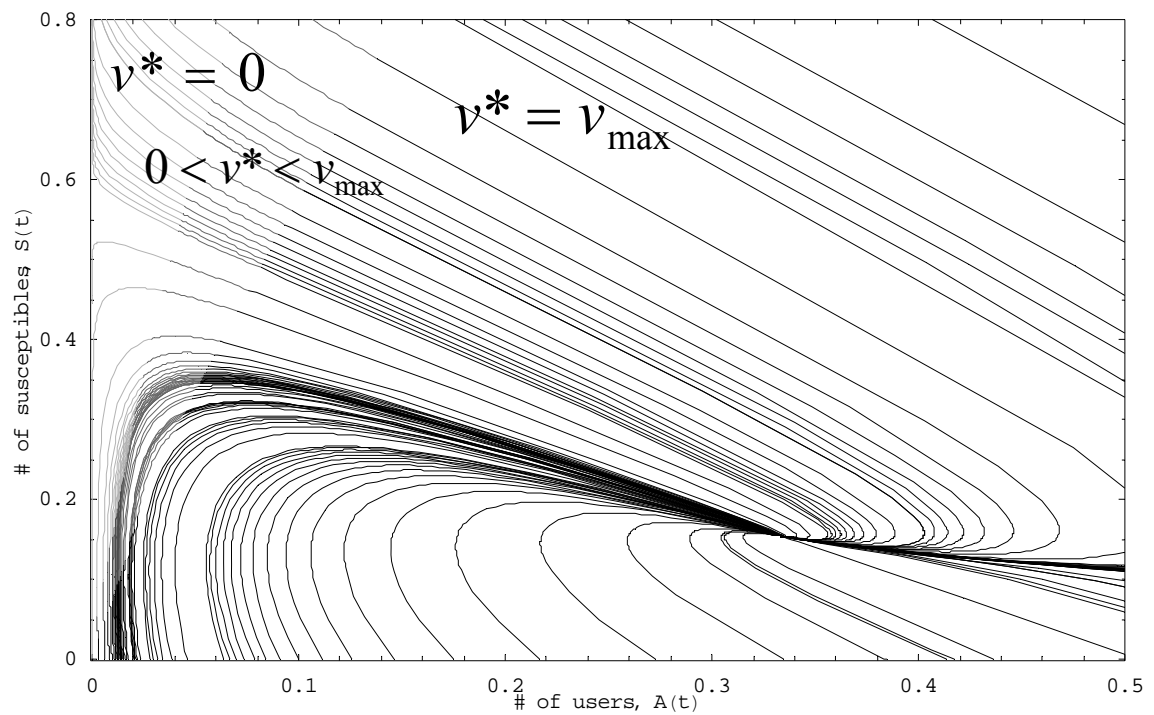
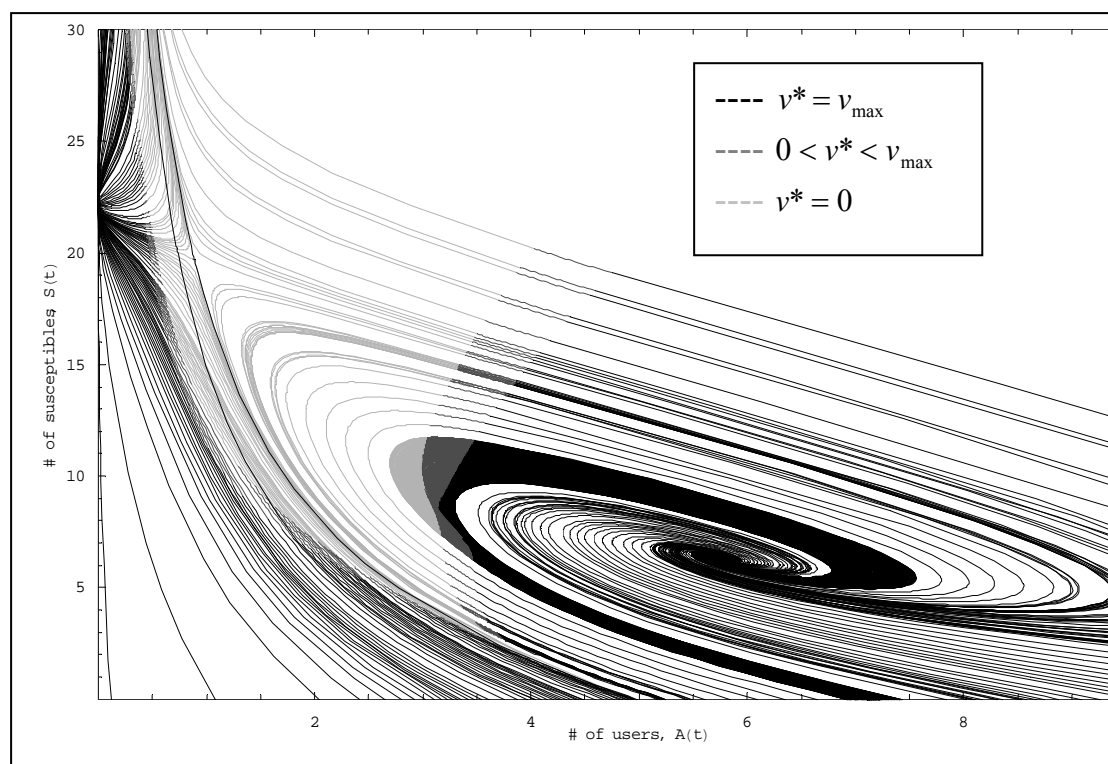


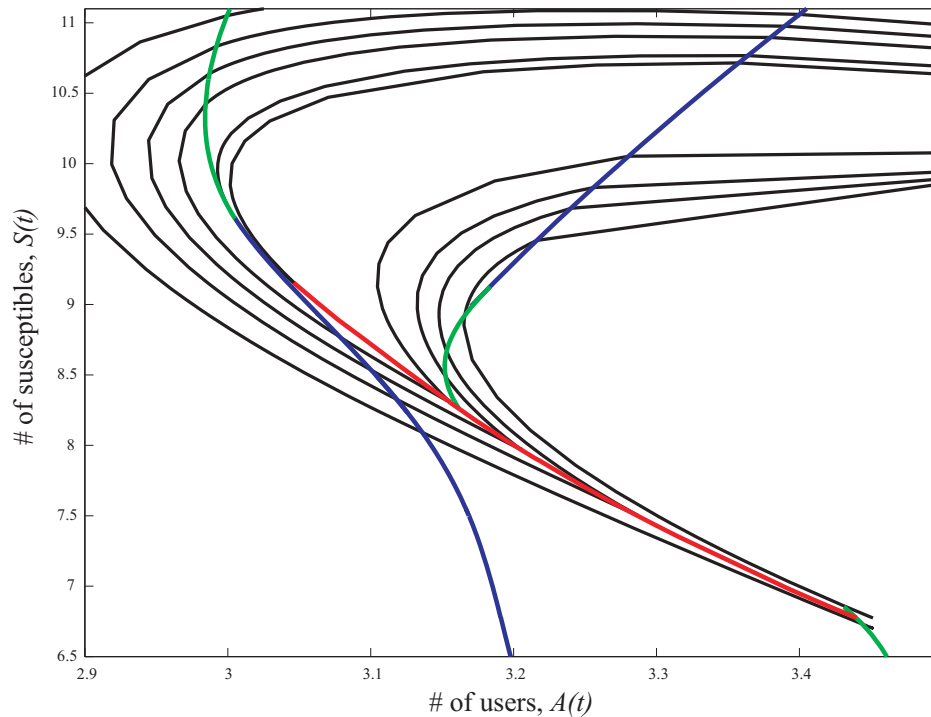
Figure 5 b) S-A-plot for optimal control strategies for the U.S.



Even in the optimal control formulation, the SA model parameterized for US cocaine retains its “tipping points” but they are now more properly called indifference curves or DNSS curves and some have a quite different character. Instead of only separating trajectories that approach *different* long-run steady states, there now are also indifference points separating different trajectories that approach the *same* long run steady state, but in very different ways.

In particular, we encounter a curve in the region of approximately $2.9 < A < 3.45$ and $7 < S < 9.5$, where for initial points on that curve two different optimal paths exist. They involve distinct optimal policies, but induce the same objective function cost in the long run and the same steady state will be achieved. In Figure 6, the red curve identifies the set of points that exhibit this feature.

Figure 6: The red curve identifies a so-called curve of points of indifference in the U.S. parameterization



6.0 Discussion

Drug abuse can do enormous damage to users and to society more generally, so it is natural to ask whether drug use can be made safer. To varying degrees the answer is, “Yes, drugs can be made safer, although by no means safe.” Some countries (e.g., Australia) have embraced such “harm reduction” interventions. Others (e.g., the US) eschew harm reduction, in no small part because opponents worry that making drugs safer might encourage greater use.

Sensitive to this possibility, MacCoun (1998) reviews evidence concerning behavioral “risk compensation” in a variety of domains and suggests that people do often participate in an activity more frequently when it is safer, but the increases are smaller, proportionately, than the reductions in harm, so total harm is generally reduced when an activity is made less harmful.

That the direct behavioral response is less than proportionate need not, however, imply that the total long-run change in use is less than proportionate if there is feedback in the system governing the dynamic evolution of use. It is generally accepted that social-interactions and/or market effects can generate such feedback in trajectories of drug prevalence, so it worth moving beyond a simple, static

conceptualization that total harm = use * average harmfulness, to embed harm reduction interventions within a dynamic model of the evolution of drug use.

The results above are a proof by example that in a system with feedback, reductions in harm that have a less than proportionate direct effect on initiation can still have a much more than proportionate effect on long-run levels of use. It is noteworthy that the epidemic model employed was an exceedingly simple one. It has just two states (users and non-users who are susceptible to initiation) and only one non-linearity – namely that initiation arises from the mixing of users and susceptible non-users, with the probability that such interactions generate a new initiation possibly depending on whether the drug is widely or not so widely used. Hence, one does not have to concoct a particularly exotic set of non-linear feedbacks for dynamics to matter.

A particular mechanism that emerged from this simple model of drug use dynamics was a tipping point separating two stable equilibria, one with low-levels of use and the other with a much higher level of use. If the system is initially close to the tipping point but on the side leading to low-levels of use, implementing harm reduction might shift the tipping point, so that the current level of use is then on the side of the tipping point leading use to grow toward the high-level equilibrium.

This mechanism emerged in the model parameterized for US cocaine principally because the parameters suggest that random interactions between users and susceptible non-users were *more* likely to lead to initiation when cocaine use was common. In contrast, the parameterization for Australian IDU suggests that the likelihood such interactions lead to an initiation is *decreasing* in the current prevalence of use. This difference accounts in no small measure for the model result that harm reduction is always a good idea for Australian IDU, but there is a (relatively narrow) range of initial conditions for which harm reduction increases not just use but also total harm for the model parameterized for US cocaine use. Those troublesome initial conditions involved levels of use far below current levels, so if the model were to be interpreted literally, it would suggest that implementing harm reduction for US cocaine today would reduce the present value of total future harm.

The models are highly stylized and parameterizations tenuous, so it is important not to put too much stock in those specific recommendations. However, many models of drug epidemics have tipping points and for various reasons (e.g.,

“enforcement swamping”), not just because of the specific mechanism in play here. So the caution about harm reduction is a more general one.

If we want to derive a policy prescription, it becomes “Harm reduction is least likely to be problematic when one is confident the drug problem is not near a point where modest perturbations favoring greater use can be multiplied into large changes in use.”

At the outset of this research, we expected the bottom line recommendation to be, “Harm reduction can safely be implemented late in an epidemic, when use has stabilized near endemic levels.” However, we identified – without great effort – examples where feedback can make harm reduction counter-productive even when use had already stabilized at endemic levels.

A more general implication is that the bitterly opposing sides of the harm reduction debate may both have valid points. Common ground – or at least a more productive characterization of differences – might come from both sides articulating more carefully their presumptions concerning the dynamics of drug use. Then both sides could couch their recommendations in conditional language, conditional on those presumptions. For example, a harm reduction advocate might say, “Harm reduction is worth considering for this drug in this country because of the following evidence concerning its dynamics” rather than saying “Harm reduction is unequivocally and universally the best policy”.

References

- Agar, M. H., & Wilson, D. (2002). Drugmart: Heroin epidemics as complex adaptive systems. *Complexity*, 7(5), 44-52.
- Anthony J.C., Warner L.A., & Kessler R.C. (1994). Comparative epidemiology of dependence on tobacco, alcohol, controlled substances, and inhalants: Basic findings from the National Comorbidity Survey. *Experimental and Clinical Psychopharmacology* 2:244-268.
- Bachman, J.G., Johnston, L.D., & O'Malley, P.M. (1998). Explaining the recent increases in students' marijuana use: The impacts of perceived risks and disapproval from 1976 through 1996. *American Journal of Public Health* 88:887-892.
- Bass FM. A new product growth model for consumer durables. *Management Science* 1969;15; 215-227.
- Baveja A, Batta R, Caulkins JP, Karwan MH. Modeling the response of illicit drug markets to local enforcement. *Socio-Economic Planning Sciences* 1993;27; 73-89.
- Behrens DA, Caulkins JP, Tragler G, Haunschmied JL, Feichtinger G. A dynamical model of drug initiation: Implications for treatment and drug control. *Mathematical BioSciences* 1999;159; 1-20.
- Behrens DA, Caulkins JP, Tragler F, Haunschmied J, Feichtinger G. Optimal control of drug epidemics: Prevent and treat – but not at the same time. *Management Science* 2000;46; 333-347.
- Bretteville-Jensen, A-L. (2006) “Drug Demand-Initiation, Continuation and Quitting” *De Economist* **154 (4)** 491-516.
- Brill, H., & Hirose, T. (1969). The rise and fall of a methamphetamine epidemic: Japan 1945-55, *Seminars in Psychiatry*, 1(2), 179-194.
- Caulkins JP. Local drug markets' Response to focused police enforcement. *Operations Research* 1993;41;848-863.
- Caulkins J. Models Pertaining to How Drug Policy Should Vary Over the Course of an Epidemic Cycle. In: Lindgren B, Grossman M. (Eds), *Substance use: individual behavior, social interactions, markets, and politics, advances in*

- health economics and health services research, vol. 16. Elsevier, 2005. p. 407-439.
- Caulkins JP, Dietze P, Ritter A. Dynamic compartmental model of trends in Australian drug use. *Healthcare Management Science* 2007;10; DOI: 10.1007/s10729-007-9012-0.
- Caulkins, Jonathan P., Gustav Feichtinger, Gernot Tragler, Dagmar Wallner (in submission). When In A Drug Epidemic Should the Policy Objective Switch from Use Reduction to Harm Reduction? Submitted to the *European Journal of Operations Research*.
- Caulkins JP, Pacula R, Paddock S, Chiesa J. 2002. School-based drug prevention: what kind of drug use does it prevent? MR-1459-RWJ. RAND, Santa Monica, CA.
- Caulkins JP, Reuter P. Setting goals for drug policy: harm reduction or use reduction *Addiction* 1997; 92;1143-1150.
- Collins DJ, Lapsley HM. Counting the cost: Estimates of the social costs of drug abuse in 1998-99. Canberra: Commonwealth of Australia; 2002.
- Coomber, R (2006) *Pusher Myths: Re-Situating the Drug Dealer*. London: Free Association Books.
- Dave, Dhaval. (2006) The effects of cocaine and heroin price on drug-related emergency department visits. *J. Health Economics* **25**, 311-333.
- Dave, Dhaval (2008) Illicit drug use among arrestees, prices and policy. *Journal of Urban Economics*, 63:694-714.
- Desimone J; Farrelly MC. Price and enforcement effects on cocaine and marijuana demand. *Economic Inquiry* 41(1): 98-115, 2003.
- Drummond M. Return on Investment in Needle and Syringe Exchange Programs in Australia. Commonwealth Department of Health and Ageing; Canberra: Commonwealth of Australia; 2004.
- Farrelly, M.C., J.W. Bray, G. A. Zarkin, B. W. Wendling, and R.L. Pacula, "The Effects of Prices and Policies on the Demand for Marijuana: Evidence from the National Household Surveys on Drug Abuse" National Bureau of Economic Research Working Paper number 6940, 1999.
- Ferrence R. Diffusion theory and drug use. *Addiction* 2001, 96: 165-173.

- Gable RS. Comparison of acute lethal toxicity of commonly abuse psychoactive substances. *Addiction* 2004; 99;686-696.
- Golub, A. and B.D. Johnson (1996) “The Crack Epidemic: Empirical Findings Support a Hypothesized Diffusion of Innovation Process,” *Socio-Economic Planning Sciences*, **30**(3): 221-231.
- Grossman, M. (2005) “Individual Behaviors and Substance Use: The Role of Price” in Lindgren, B. and M. Grossman (eds.) *Substance Use: Individual Behaviors, Social Interactions, Markets and Politics* Advances in Health Economics and Health Services Resesarch Vol. 16, Elsevier, Amsterdam.
- Harwood H, Fountain D, Livermore G. The economic costs of alcohol and drug abuse in the United States, 1992. Rockville, MD: National Institutes on Drug Abuse; 1998.
- Hunt, L.G., and C.D. Chambers. 1976. *The Heroin Epidemic: A Study of Heroin Use in the U.S., 1965-1975 (Part II)*. Holliswood, New York: Spectrum.
- Kirby KN, Petry NM. Heroin and cocaine abusers have higher discount rates for delayed rewards than alcoholics or non-drug-using controls. *Addiction* 2004; 99; 461-471.
- Kleiman MAR 1988. Crackdowns: The effects of intensive enforcement on retail heroin dealing. in *Street-Level Drug Enforcement: Examining the Issues*. In: Chaiken MR (ed), National Institute of Justice, Washington DC.
- Kleiman MAR. Enforcement swamping: A positive-feedback mechanism in rates of illicit activity. *Mathematical and Computer Modeling* 1993;17; 65-75.
- Kleiman MAR. “Economic cost” measurements, damage minimization and drug abuse control policy. *Addiction* 1999; 94; 638-634.
- MacCoun RJ, Reuter P, Schelling T. Assessing alternative drug control regimes. *Journal of Policy Analysis and Management* 1996, 15, 330-352.
- MacCoun RJ. Toward a psychology of harm reduction. *American Psychologist* 1998; 53; 1199-1208.
- McCaffrey BR. Legalization would be the wrong direction. *Los Angeles Times*, July 27, 1998.
- Moore MH. *Buy and bust: The effective regulation of an illicit market in heroin*. Lexington: MA; Lexington Books. 1977.

- Moore TJ. (2005). Monograph No 01: What is Australia's "Drug Budget"? The policy mix of illicit drug-related government spending in Australia. DPMP Monograph Series. Fitzroy: Turning Point Alcohol and Drug Centre.
- Moore TJ, Caulkins JP. How cost-of-illness studies can be made more useful for illicit drug policy analysis. *Applied Health Economics and Health Policy* 2006;5;75-85.
- Musto D. *The American disease*, 2nd Edition. New Haven: CT; Yale University Press. 1987.
- Nordt C, Stohler R. Incidence of heroin use in Zurich, Switzerland: A treatment case register analysis. *The Lancet* 2006; 367;1830-1834.
- Noymer A. The transmission and persistence of 'Urban Legends': Sociological application of age-structured epidemic models. *Journal of Mathematical Sociology* 2001, 25;299-323.
- Pacula, R.L, M. Grossman, F.J. Chaloupka, P.M. O'Malley, and M.C. Farrelly 2001. Marijuana and youth. In *Risky Behavior Among Youths: An Economic Analysis*, ed. J. Gruber. Chicago: University of Chicago Press, 271-326.
- Reuter P. Are calculations of the economic costs of drug abuse either possible or useful? *Addiction* 1999; 94;635-637.
- Rigter H. What drug policies cost: Drug policy spending in the Netherlands in 2003. *Addiction* 2006;101;323-329.
- Riley KJ. Crack, powder cocaine, and heroin: drug purchase and use patterns in six U.S. cities. National Institute of Justice, Washington: DC. 1997.
- Rossi, Carla. 2001. "A Mover-Stayer Type Model for Epidemics of Problematic Drug Use," *Bulletin on Narcotics*, Vol. 53, No. 1, pp.39-64.
- Saffer, H. and F.J. Chaloupka 1999. The demand for illicit drugs. *Economic Inquiry* 37(3): 401-411.
- Shapiro, Carl, and Hal Varian (1998) *Information Rules: A Strategic Guide to the Network Economy*. Harvard Business School Press.
- Single E, Collins D, Easton B, et al. *International Guidelines for Estimating the Costs of Substance Abuse: Second edition*. Geneva: World Health Organization, 2003.
- Tragler G, Caulkins JP, Feichtinger G. Optimal dynamic allocation of treatment and enforcement in illicit drug control. *Operations Research* 2001;49;352-362.

- United Nations Office on Drugs and Crime. 2005 World Drug Report. Oxford University Press, 2005.
- Wilde, Gerald J.S. (1994) *Target Risk*. PDE Publications. Available online at <http://psyc.queensu.ca/target/index.html>.
- Williams, J. (2004), 'The Effects of Price and Policy on Marijuana Use: What can be Learned from the Australian Experience?,' *Health Economics*, 13, pp. 123–137.
- Williams, J., R.L. Pacula, F.J. Chaloupka and H. Wechsler (2006), 'College Students' Use of Cocaine,' *Substance and Misuse*, 41, pp. 489–509.
- Winkler D, Caulkins JP, Behrens D, Tragler G. Estimating the relative efficiency of various forms of prevention at different stages of a drug epidemic. *Socio-Economic Planning Sciences* 2004;38;43-56.