RESPECT THE UNSTABLE: DELAYS AND SATURATION IN CONTACT TRACING FOR DISEASE CONTROL *

RICHARD PATES†, ANDRES FERRAGUT‡, ELIJAH PIVO§, PENGCHENG YOU¶, FERNANDO PAGANINI‡, AND ENRIQUE MALLADA¶

Abstract. Motivated by the novel coronavirus disease (COVID-19) pandemic, this paper aims to apply Gunter Stein's cautionary message of *respecting the unstable* to the problem of controlling the spread of an infectious disease. With this goal, we study the effect that delays and capacity constraints have in the test, trace and isolate (TeTrIs) process, and how they impact its ability to prevent exponential disease spread. Our analysis highlights the critical importance of speed and scale in the TeTrIs process. Precisely, ensuring that the delay in the TeTrIs process is much smaller than the doubling time of the disease spread is necessary for achieving acceptable performance. Similarly, limited TeTrIs capacity introduces a threshold on the size of an outbreak beyond which the disease spreads almost like the uncontrolled case. Along the way, we provide numerical illustrations to highlight these points.

Key words. feedback control, stabilization, epidemic spread, COVID-19

AMS subject classifications. 93D15, 93D09, 93D20, 92D25, 92D30

1. Introduction.

The opening lines of Gunter Stein's classic paper *Respect the Unstable* [24], published 13 years after his inaugural Bode Lecture of the same name, read:

> "The practical, physical (and sometimes dangerous) consequences of control must be respected, and the underlying principles must be clearly and well taught."

The message to the control engineer and researcher is clear. Not only must the many benefits of feedback be understood (pedagogically, mathematically, and in practice), but also its limitations. The principle of feedback is after all inherently about trade-offs, constrained by conservation laws just as fundamental as any law of physics. Whilst these 'laws of feedback' apply to the control of all systems, Gunter Stein gave special attention to unstable systems for three main reasons:

1. Unstable systems are fundamentally, and quantifiably, more difficult to control than stable ones.
2. Controllers for unstable systems are operationally critical.
3. Closed-loop systems with unstable components are only locally stable.

In this paper we aim to revisit these points from the perspective of designing contact tracing policies to mitigate the spread of disease throughout a population.

1.1. Control of Disease Spread.

The control of disease spread is not the traditional hunting ground of the control engineers, so a degree of caution from our community is perhaps of even greater relevance than normal. That said, controlling the spread of a disease has many of the elements of the most challenging control

*Submitted to the editors Nov. 1st, 2020.

Funding: This work was funded by NSF through grants CAREER 1752362 and TRIPODS 1934979, and Johns Hopkins University Discovery Award.

†Lund University, Lund, Sweden. (richard.pates@control.lth.se).

‡Universidad ORT, Montevideo, Uruguay, (ferragut@ort.edu.uy,paganini@ort.edu.uy)

§Massachusetts Institute of Technology, Boston, MA USA, (epivo@mit.edu)

¶Johns Hopkins University, Baltimore MD, USA (pcyou@jhu.edu,mallada@jhu.edu)

problems. Accurate models of the spread of a highly infectious disease are at best controversial, but certainly unstable (at least in a population with high susceptibility to the disease). The mechanisms for identifying infectious members of the population may be subject to significant delays and inaccuracies, compromising the quality of the available information for performing feedback. And finally, the options for mitigating the spread can be blunt, unpredictable, and subject to severe capacity constraints.

Since emerging in late 2019, the novel coronavirus disease (COVID-19) pandemic has made abundantly clear the effect that these challenges have on mitigating disease spread. At the time of writing (Oct. 2020), COVID-19 had reached a significant global spread (45 M documented cases) [7] and vaccines were not yet available; this meant that the primary public health tools available to limit the spread were non-pharmaceutical interventions (NPIs), such as social distancing and contact tracing [11]. Many NPIs can be understood in terms of feedback control, and as such abide by the fundamental 'laws of feedback' that Gunter Stein referred to. This work illustrates the impact of these limitations, placing a particular emphasis on the role of delays and saturation. We focus on contact tracing as it exhibits several of the features described above.

1.2. Contact Tracing. Contact tracing is the process of testing, tracing and isolating people known to have been in close proximity with infected individuals. All three of these steps are essential, so for this reason contact tracing is also referred to by the acronym TeTrIs. This intervention can disrupt chains of infection to slow and potentially end the spread of an infectious disease. It has been employed in the control of sexually transmitted infections [6, 12, 19], in limiting the severe acute respiratory syndrome (SARS) epidemic [5] and at an unprecedented scale in the COVID-19 pandemic [23, 1].

The execution of TeTrIs varies significantly from region to region, and is rapidly evolving. Regardless of the specifics, two key characteristics contribute to the success of TeTrIs. The first is the delay between the moment an individual becomes infected and the moment that individual becomes isolated from the rest of the population. A larger delay allows the infected individual to infect more people. The second is the capacity of the TeTrIs program. We think of this capacity as the number of active cases the TeTrIs program can process at once without the delay growing significantly. These characteristics are determined by the structure of the TeTrIs program. But more practically, achieving sufficient performance in these characteristics must be used to determine the structure of the TeTrIs program. Thus, in this paper we seek to characterize sufficient delays and capacity of a TeTrIs program to successfully control the spread of an infectious disease.

The effects of these characteristics have been studied in the past. Many works analyze the impacts of contact tracing using computer simulations [18, 10]. Mathematical analysis of TeTrIs has typically relied on two methodologies. In the first, an ordinary differential equation (ODE) models spread over a certain fixed contact graph [9, 14]. In the second, the impact of TeTrIs is modeled as a branching process [21, 20].

1.3. Contributions of this Work. In this work, we take a control theoretic perspective on the impacts of delays and saturation. These two phenomenon have been widely studied in the control systems field. We provide two rules of thumb for the requisite speed and capacity of a TeTrIs system. First by analyzing the system sensitivity function, we show that delays of even just one quarter of the doubling time of the disease may suffice to overwhelm a TeTrIs system. For infectious diseases

like COVID-19, the optimistic allowable delay to control their initial outbreak is approximately 1 day. Another implication of the analysis points to the importance of effective isolation. If we fail to isolate two thirds of the cases, such a system may not even be stabilising without delay. Second, we model the contact tracing process and show that the saturation of its limited capacity may disable an otherwise efficacious TeTrIs system. With saturation, we identify a threshold behavior of disease spread that implies stability regions beyond capacity and potentially significant degradation of performance.

The paper is structured as follows. First, we discuss the effects of delay on the efficacy contact tracing. We introduce contact tracing as a feedback loop on the classic SIR model. We derive an upper bound on delay to prevent exponential disease spread in this setting. Then, we generalize this analysis from the SIR model to general LTI and nonlinear system models with an exponentially unstable mode. This demonstrates that these limitations are fundamental, rather than an artifact of particular modelling choices. Second, we discuss the effects of saturation on the efficacy of contact tracing. We introduce two compartmental models that respectively capture the contact tracing efforts devoted to infected and uninfected populations and introduce the saturation effects of tracing capacity. Reduced stability regions are observed based on a nonlinear threshold analysis.

Notation. Transfer functions of linear-time-invariant (LTI) systems will be denoted with bold face letters. For example $\mathbf{G}(s) = 1/(s+1)$ is the transfer function from u to x for the system $\frac{dx}{dt} = -x + u$, and $\mathbf{G}(s) = \exp(-sT)$ the transfer function for the delay $x(t) = u(t-T)$. The set of all proper real rational transfer functions, i.e. functions of the form

$$\mathbf{G}(s) = \frac{a_0 s^n + a_1 s^{n-1} + \ldots + a_n}{s^n + b_1 s^{n-1} + \ldots + b_n}, a_i \in \mathbb{R}, b_k \in \mathbb{R}$$

will be denoted by $\mathscr{R}$. The H-infinity norm of a transfer function $\mathbf{G}$ is defined as

$$\|\mathbf{G}\|_\infty \coloneqq \sup\{|\mathbf{G}(s)| : s \in \mathbb{C}, \mathrm{Re}(s) > 0\}.$$

The H-infinity norm is a central notion in the robust performance of control systems, see for example [8, §2] for an introduction.

2. Contact tracing: The Need for Speed.

The basic rationale behind TeTrIs is simple. Disease spreads through the contact between infectious and susceptible members of a population. So by rapidly isolating infectious individuals as soon as they are detected, as well as everyone they've recently contacted (who may now be infectious themselves), it may be possible to shut off all the routes of spread, and stop an outbreak in its tracks. But how accurate does the testing need to be to ensure that enough cases are traced? And how fast must the system be to halt an outbreak before it becomes an epidemic?

In this section we will explore these questions from the control-theoretic perspective, with a particular focus on feedback based fundamental limitations. TeTrIs is a feedback process, in which infectious people are isolated in response to measurements about a population. Therefore, TeTrIs is subject to conservation laws and performance limitations (see [24, 2] for an introduction). We will discuss the consequences of these, placing a particular focus on the following inequality:

$$\|\mathbf{S}\|_\infty \geq 2^{\frac{T_{\text{delay}}}{T_{\text{doubling}}}}. \tag{2.1}$$

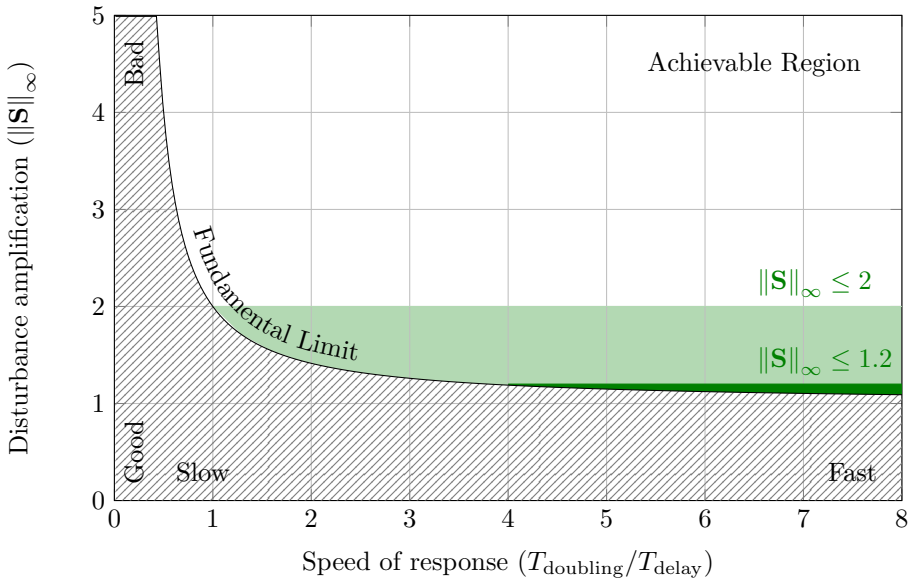


FIG. 1. *Trade-off between disturbance amplification and time delay when controlling an unstable system. Typically* $\|\mathbf{S}\|_\infty$ *less than 1.2–2 is necessary for good performance.*

The precise meanings of all these terms will be made clear when it is derived in Subsection 2.2, but here $\mathbf{S}$ is the sensitivity function (in the usual control theoretic sense), T_{doubling} the doubling time of the unstable process[1], and T_{delay} the sum of delays in the feedback loop. This inequality imposes a fundamental limit on the size of the sensitivity function, and shows that when very unstable processes (smaller doubling times) are controlled subject to large delays, the sensitivity function will always be large. This is illustrated in Fig. 1. Since the sensitivity function determines how disturbances are amplified and attenuated, (2.1) demonstrates that in such systems, bad performance is inevitable. Indeed the conventional wisdom is that a value of $\|\mathbf{S}\|_\infty$ less than 1.2–2 is a prerequisite for acceptable performance (see e.g. [3, 8]). The size of $\|\mathbf{S}\|_\infty$ is also intimately related to many other measures of performance and robustness, such as gain and phase margins [3, §7.2].

Equation (2.1) gives the implication

$$T_{\text{delay}} > T_{\text{doubling}} \log_2 k_{\text{perf}} \quad \Longrightarrow \quad \|\mathbf{S}\|_\infty > k_{\text{perf}}.$$

The consequences of this inequality are quite striking in the context of controlling disease spread using TeTrIs. For example, it shows that given a disease with a doubling time of 8 days, if the delays between becoming infectious and being isolated are greater than 2 days, then $\|\mathbf{S}\|_\infty > 1.2$ (picking the more conservative target might be advisable when trying to control a highly uncertain system such as disease spread). This bound holds even under extremely optimistic assumptions about the implementation of contact tracing. Specific implementations can certainly be worse!

What makes the bound useful is that it provides direct insight into our original questions. For example, if we set a target of $\|\mathbf{S}\|_\infty \leq 1.2$, the system set up to

[1] Here $T_{\text{doubling}} := \frac{\ln 2}{p}$, where $p > 0$ is the location of the unstable pole.

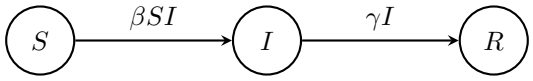
conduct contact tracing must be at least four times faster than the doubling time of the disease:

$$\|\mathbf{S}\|_\infty \leq 1.2 \quad \Longrightarrow \quad 4T_{\text{delay}} \leq T_{\text{doubling}}.$$

Slower implementations are guaranteed to fail this objective, and as a result be more vulnerable to disturbances (e.g. failing to identify an infectious person could result in a large number of new infections). It is interesting to note that the same rule of thumb based on more ad-hoc arguments can be found in [4, §III.B-4)]. Inequalities such as (2.1) provide further evidence for the necessity of a fast TeTrIs system.

2.1. Understanding the Issue. In this section we will demonstrate the fundamental limitation discussed above from the perspective of a simple model of contact tracing. This will allow us to put these abstract ideas in a more concrete setting, so as to better understand them. Studying a simple model will also allow us to derive specialised analysis tools along the way that can provide additional insight. In what follows we will first outline a simple SIR-based model for contact tracing, before illustrating the fundamental limitations through simulations and additional theoretical tools.

2.1.1. An SIR-based Model for Disease Control with TeTrIs. The so called SIR model is one of the simplest and most widely used models of disease spread [16]. It is centred around three compartments - $S(t)$, $I(t)$ and $R(t)$ - which specify the proportion of the population that are susceptible, infectious, and recovered at time t. So if $S(0) = 1$, then at time $t = 0$ the entire population is susceptible to the disease, or if $R(1) = 0.5$ then half the population has recovered (or died) at time $t = 1$. The population shifts between these compartments over time according to two rates, which model the effect of the infectious population mixing with the susceptible population and transferring the disease, and the infectious population recovering, respectively. This can be visualised on a graph with a node for each compartment, and a directed edge specifying the transition rates between them:



Here β is a mixing parameter, specifying the average number of 'significant' (those that could result in the transmission of the disease) interactions that each individual has per unit time. Each infectious person then has an average of βS such events with the susceptible population, resulting in βSI new infections per unit time. The second rate is justified by saying that on average it takes $1/\gamma$ units of time for an infectious person to recover, which corresponds to members of the I compartment being transferred to the R compartment with rate γI.

When written as a set of differential-algebraic equations, the SIR model is

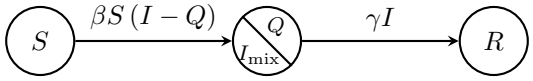
$$\frac{d}{dt}\begin{bmatrix} S \\ I \\ R \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix} \beta SI + \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix} \gamma I, \quad 1 = S + I + R. \tag{2.2}$$

Of central importance in the study of the SIR model (and disease spread in general) is the so-called basic reproduction number R_0. R_0 is defined to be the number of secondary infections caused by a single primary infection in a population in which everyone is susceptible to the disease. Consequently if $R_0 > 1$ a small outbreak will grow, whereas if $R_0 < 1$ it will not. For the SIR model, $R_0 = \beta/\gamma$. This is closely

related to notions of stability and doubling times. For the SIR model

$$T_{\text{doubling}} = \frac{\ln 2}{\beta - \gamma} = \frac{\ln 2/\beta}{1 - 1/R_0}. \tag{2.3}$$

The SIR model describes the process of disease spread, but not the impact of TeTrIs. To model this, we first split the infectious population into two groups Q and I_{mix}, where Q corresponds to the subpopulation that has been quarantined, and I_{mix} the remainder of the infectious population. We can incorporate the effect of quarantining, by modifying the rate between the susceptible and infectious population as shown below. The rationale here is that after taking quarantining into account there should be $\beta S I_{\text{mix}}$ new infections per unit time, and that $I_{\text{mix}} = I - Q$.



The effect of this change is to slightly modify the original SIR equation in (2.2):

$$\frac{d}{dt}\begin{bmatrix} S \\ I \\ R \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix} \beta S (I - Q) + \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix} \gamma I, \quad 1 = S + I + R. \tag{2.4}$$

All that remains is to close the loop, and specify how the number of people who are quarantined at time t depends on the contact tracing. For simplicity, we propose to model this process through the equation

$$Q(t) = \alpha e^{-\gamma T_{\text{delay}}} I(t - T_{\text{delay}}), \tag{2.5}$$

where $1 \geq \alpha \geq 0$ and $T_{\text{delay}} \geq 0$. In words this equation says that we are able to test, trace and isolate a proportion α of those that were infectious T_{delay} days ago[2]. Together (2.4) and (2.5) constitute a simple model for understanding how TeTrIs can be used to control disease spread.

2.1.2. Analysis of the Simple Model. Before performing a theoretical analysis of the model, it is instructive to run some simulations. The evolution of the infectious population after an outbreak affecting 0.01% of the population is shown in Fig. 2 for a range of different values of the time delay. The simulation parameters for this figure are:

- $\alpha = 0.8$, meaning that 80% of cases are tested, traced and isolated.
- $\gamma = 0.1$, meaning the disease has an average recovery time of 10 days.
- $\beta = 0.3$, giving the disease a basic reproduction number of 3.

The first thing to note is that if the delay is short, the outbreak is contained and no epidemic ensues. It is also interesting to see the degradation in behaviour as the delay increases. By the time T_{delay} is 5 days, an epidemic not dissimilar to that without TeTrIs occurs. Even more strikingly though is that by the time T_{delay} is just 2 days, the initial outbreak sees a tenfold increase before it is brought under control. This relatively short delay has seemingly brought TeTrIs to the verge of instability. When you consider that there may be several simultaneous outbreaks, or capacity constraints on how many people that can be tested-and-traced, it is clear that short delays may already be enough to overwhelm a TeTrIs system.

[2]We need to include the proportional constant $e^{-\gamma T_{\text{delay}}}$ since over those T_{delay} days, $(1 - e^{-\gamma T_{\text{delay}}})$ of those that were infectious will have gone on to recover.

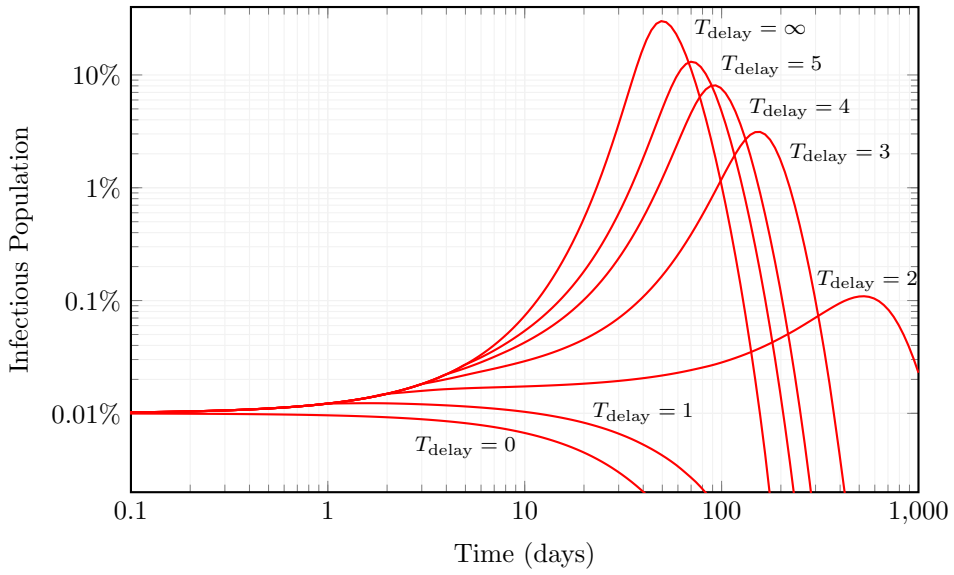


FIG. 2. *Simulation of (2.4) and (2.5) for a range of values of* T_{delay}.

A natural first question is, "*Are these results in line with the fundamental limitation discussed at the beginning of this section?*". A simple calculation shows that at the start of the outbreak, the doubling time of the disease equals

$$T_{\text{doubling}} = \frac{\ln 2}{\beta - \gamma} \approx 3.5 \text{ days}.$$

Therefore, to achieve $\|\mathbf{S}\|_\infty \leq 1.2$, it is necessary that $T_{\text{delay}} \leq 0.9$ days. This seems to be in good agreement with the simulation, where the case with a one day delay is well controlled, with a rapid decline in performance soon after. In fact, given the simple nature of the model in (2.4) and (2.5) a more detailed analysis is possible. The following theorem characterises the stability of the linearisation of the model about the disease free equilibrium in terms of the system parameters. An intuitive explanation of this stability criterion is given at the end of the section.

THEOREM 2.1. *The linearisation of the model in (2.4) and (2.5) is stable*[3] *about the point* $(I, R, Q) = (0, 0, 0)$ *if and only if*

$$T_{\text{delay}} < \frac{1}{\gamma} \ln\left(\frac{\alpha\beta}{\beta - \gamma}\right). \tag{2.6}$$

Proof. See Appendix A. ∎

Remark 2.2. While any point with $I = 0$ (no infected people) is an equilibrium of (2.4) and (2.5), we focus on the point $(I, R, Q) = (0, 0, 0)$ for two reasons. Firstly, this equilibrium corresponds to the initial phase of the pandemic ($S = 1$) and exhibits the largest unstable growth, thus serving as a natural benchmark for stabilization purposes. Secondly, it is also the most desirable equilibrium from a public health

[3] In the sense that $I(t) \to 0$ in response to a small perturbation about the initial condition $(I(t), R(t), Q(t)) = (0, 0, 0)$ for $t \leq 0$.

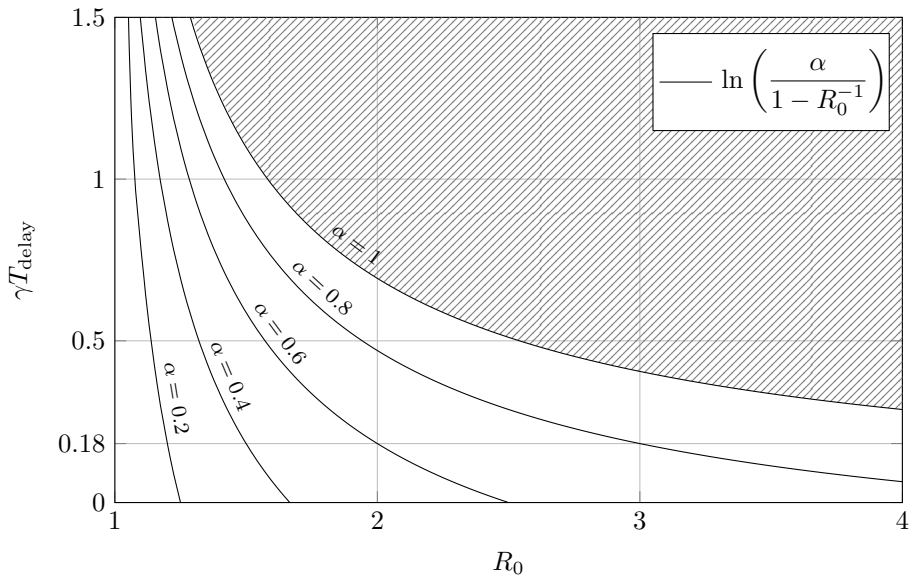


FIG. 3. *Illustration of the stability boundary in Theorem 2.1. The model of TeTrIs is stabilising if and only if* $(R_0, \gamma T_{\text{delay}})$ *lies below the corresponding* α *curve. For example, if* $\alpha = 0.8$ *and* $R_0 = 3$*, the model is stable if and only if* $\gamma T_{\text{delay}} < 0.18$

perspective and of high practical value. Indeed, many countries have achieved initial control of the COVID-19 pandemic through TeTrIs sustaining levels of infections of several order of magnitude lower than its population. For example, Uruguay, the home country of several authors of this work, sustained levels of active infections in at most hundreds for several months, over a population of approximately 3.5 M.

In order to interpret the meaning of Theorem 2.1, it helps to rearrange the bound a little:

$$\gamma T_{\text{delay}} < \ln\left(\frac{\alpha\beta}{\beta-\gamma}\right) = \ln\left(\frac{\alpha}{1-1/R_0}\right).$$

The specific trade-off between parameters and delay implied by the above is shown in Fig. 3. This figure can be used to quickly assess the amount of delay that can be tolerated before instability occurs. For example, in the simulations we used a model with $R_0 = 3$ and $\gamma = 0.1$, with feedback parameter $\alpha = 0.8$. Therefore, from the figure we see that we require

$$T_{\text{delay}}\gamma < 0.18 \quad \Longrightarrow \quad T_{\text{delay}} < 1.8 \text{ days}$$

for the policy to be stabilising. This captures precisely the behaviour we saw in the simulation, where $T_{\text{delay}} = 2$ seemed to be right on the cusp of instability. We also see the importance of tracing enough cases. By the time $\alpha < 1 - R_0^{-1} = 2/3$, that is, we only detect and isolate at most 66% of the cases, the policy isn't even stabilising with $T_{\text{delay}} = 0$.

The stability criterion in Theorem 2.1 also has a nice interpretation through the effective reproduction number R_e. Suppose that α in (2.5) is the probability that an infectious individual is detected and isolated. The amount of time T that each

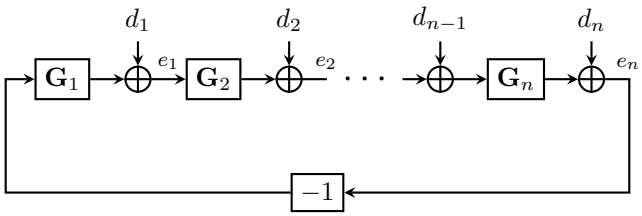


FIG. 4. *Feedback interconnection in (2.7).*

infectious person is mixing with the susceptible population is then a random variable

$$T = \begin{cases} T_r & \text{w.p.} \quad 1-\alpha \\ \min\{T_{\text{delay}}, T_r\} & \text{w.p.} \quad \alpha. \end{cases}$$

In the above $T_r \sim \text{Exp}(\gamma)$ is the time it takes the given person to recover from the disease. Therefore, the expected time that each infectious person is in the mix is given by

$$\begin{aligned} \text{E}[T] &= (1-\alpha)\,\text{E}[T_r] + \alpha\text{E}[\min\{T_{\text{delay}}, T_r\}] = (1-\alpha)\frac{1}{\gamma} + \alpha\int_0^{T_{\text{delay}}} \exp(-\gamma s)\, ds \\ &= \frac{1}{\gamma}\left(1 - \alpha\exp(-\gamma T_{\text{delay}})\right). \end{aligned}$$

The effective reproduction number is then the expected number of secondary infections generated by an individual:

$$R_e = \beta\text{E}[T] = \frac{\beta}{\gamma}\left(1-\alpha\exp(-\gamma T_{\text{delay}})\right) = R_0\left(1-\alpha\exp(-\gamma T_{\text{delay}})\right).$$

The condition that $R_e < 1$, which would correspond to an outbreak dying out, is thus equivalent to

$$1 > R_0\left(1-\alpha\exp(-\gamma T_{\text{delay}})\right) \quad \Longleftrightarrow \quad T_{\text{delay}} < \frac{1}{\gamma}\ln\left(\frac{\alpha}{1-R_0^{-1}}\right),$$

which is precisely the stability condition from Theorem 2.1.

2.2. Fundamental Limitations. A natural concern with the results from Subsection 2.1.2 is that they are seemingly based on a set of highly contentious modelling assumptions. For example, why use the SIR model to capture the effect of disease spread in (2.4), rather than the SEIR model or indeed any of the other more complex compartmental variants? What about other models for TeTrIs? Will the same conclusions hold if we use something more realistic than (2.5)? In this section we will demonstrate that the limitations we observed through Theorem 2.1 and the simulations of (2.4) and (2.5) are really a consequence of the interplay between instability and delay.

The main result of this section is to derive the inequality (2.1). For simplicity we will stick to the LTI case, though we will show in Appendix B that a natural analogue of (2.1) holds in the nonlinear case also. To this end, consider the feedback interconnection of n subsystems described by

$$\begin{aligned} &\hat{e}_i = \mathbf{G}_i\hat{e}_{i-1} + \hat{d}_i, \ i \in \{1,\dots,n\} \\ &\hat{e}_0 = -\hat{e}_n. \end{aligned} \tag{2.7}$$

In the above the variables $\hat{d}_i$ and $\hat{e}_i$ denote the Laplace transforms of a set of scalar disturbances and error signals, and $\mathbf{G}_i$ the transfer function of the i-th subsystem. The basic setup is illustrated in Fig. 4. This is a general framework for describing feedback systems, and many models for the control of a disease using TeTrIs can be put in this framework. For example, after linearisation about the point $(I, R, Q) = (0, 0, 0)$, the model in (2.4) and (2.5) can be captured by setting $n = 2$, and

$$\mathbf{G}_1(s) = \frac{\beta}{s - (\beta - \gamma)}, \quad \mathbf{G}_2(s) = \alpha \exp(-sT_{\text{delay}}). \tag{2.8}$$

Variants with, for example, more complicated compartmental models of disease spread can be similarly handled by substituting in the corresponding transfer function for $\mathbf{G}_1$.

The advantage of the abstract formulation in (2.8) is that it allows general properties of feedback interconnections to be studied for entire classes of models. When studying the properties of this feedback interconnection, the central objects are the sensitivity functions. These are the transfer functions from d_i to e_i, which we denote as $\mathbf{S}_i$. In the scalar LTI case, the sensitivity functions are all equal to each other and given by

$$\mathbf{S}_i = \frac{1}{1 + \mathbf{G}_1\mathbf{G}_2 \cdots \mathbf{G}_n} =: \mathbf{S}, \ i \in \{1, \ldots, n\}. \tag{2.9}$$

These functions determine how the internal signals $\hat{e}_i$ depend on the external disturbances $\hat{d}_i$. Hence the size of $\mathbf{S}$ determines how disturbances are attenuated. Indeed every single closed-loop transfer function in (2.8) contains $\mathbf{S}$ (for example the transfer function from $\hat{d}_1$ to $\hat{e}_3$ is given by $\mathbf{G}_3\mathbf{G}_2\mathbf{S}$). Given its central importance to the process of feedback, the sensitivity function has been extensively studied both in theory and in practice. Indeed the requirement that the size of $\|\mathbf{S}\|_\infty$ be less than 1.2–2 is widely used, and is arguably of more importance than criteria based on the gain margin and phase margin[4] [3, §7.2].

The following theorem shows that when the feedback loop contains a system with an unstable pole p and a time delay of T_{delay}, $\|\mathbf{S}\|_\infty \geq \exp(pT_{\text{delay}})$. This places a fundamental limit on the size of the sensitivity function. Surprisingly this result doesn't seem to be known (for example the lower bound $\|\mathbf{S}\|_\infty \geq \exp(pT_{\text{delay}}) - 1$ is presented in [3, §14.3, Table 14.1]), though the existence of such a bound is certainly implicit in the work on sensitivity optimisation from the 1980s [17, 13]. We give a simple proof based on the maximum modulus principle.

THEOREM 2.3. *If* $\mathbf{L} = \frac{\exp(-sT_{\text{delay}})}{s-p}\mathbf{H}$, *where* $T_{\text{delay}} > 0, p > 0$ *and* $\mathbf{H} \in \mathscr{R}$, *then*

$$\left\| \frac{1}{1 + \mathbf{L}} \right\|_\infty \geq \exp(pT_{\text{delay}}).$$

Proof. Let $a > 1$, and note that the Möbius transform $f(z) = (1 - az)/(a - z)$ maps the closed unit disc into the closed unit disc. This implies that given any transfer

[4] Indeed it can be shown that [3, §7.2]

$$\text{gain margin} \geq \frac{\|\mathbf{S}\|_\infty}{\|\mathbf{S}\|_\infty - 1}, \quad \text{phase margin} \geq 2\arcsin\left(\frac{1}{2\|\mathbf{S}\|_\infty}\right),$$

whereas no guarantees in the converse direction hold (positive gain and phase margins only guarantee that $\|\mathbf{S}\|_\infty < \infty$).

function $\mathbf{G}$, we have the equivalence

$$\|\mathbf{G}\|_\infty \leq 1 \quad \Longleftrightarrow \quad \|f(\mathbf{G})\|_\infty \leq 1.$$

Therefore, $\|1/(1+\mathbf{L})\|_\infty \leq a$ if and only if

$$\begin{aligned} 1 \geq \left\| f\left(\frac{1}{a}\frac{1}{1+\mathbf{L}}\right)\right\|_\infty &= \left\|\frac{a\mathbf{L}}{a^2\mathbf{L}+a^2-1}\right\|_\infty \\ &= \left\|\frac{a\mathbf{H}\exp(-sT_{\text{delay}})}{a^2\mathbf{H}\exp(-sT_{\text{delay}})+(s-p)(a^2-1)}\right\|_\infty. \end{aligned}$$

Now recall that given any transfer function $\mathbf{G}$, $\|\mathbf{G}\exp(-sT_{\text{delay}})\|_\infty = \|\mathbf{G}\|_\infty$ (delaying the input to a transfer function doesn't affect its norm). Therefore,

$$\begin{aligned} \left\|\frac{a\mathbf{H}\exp(-sT_{\text{delay}})}{a^2\mathbf{H}\exp(-sT_{\text{delay}})+(s-p)(a^2-1)}\right\|_\infty &= \left\|\frac{a\mathbf{H}}{a^2\mathbf{H}\exp(-sT_{\text{delay}})+(s-p)(a^2-1)}\right\|_\infty \\ &\geq \frac{1}{a\exp(-pT_{\text{delay}})}, \end{aligned}$$

where the inequality follows from the maximum modulus principle applied at the point $s=p$ (see e.g. [8, §6.2]). This demonstrates that $\|1/(1+\mathbf{L})\|_\infty \leq a$ only if $a \geq \exp(pT_{\text{delay}})$ as required. ∎

It is readily verified that this bound is equivalent to the inequality presented earlier in (2.1) by substituting in the relationship between p and T_{doubling}. That is, setting $p = \ln(2)/T_{\text{doubling}}$ shows that

$$\|\mathbf{S}\|_\infty \geq \exp(pT_{\text{delay}}) = 2^{\frac{T_{\text{delay}}}{T_{\text{doubling}}}}.$$

Theorem 2.3 shows that if the transfer function $\mathbf{G}_1\mathbf{G}_2\cdots\mathbf{G}_n$ (typically referred to as the return ratio) can be written in the form

$$\mathbf{G}_1\mathbf{G}_2\cdots\mathbf{G}_n = \frac{\exp(-sT_{\text{delay}})}{s-p}\mathbf{H}, \tag{2.10}$$

where $\mathbf{H}$ is any transfer function in $\mathscr{R}$, then $\|\mathbf{S}\|_\infty \geq \exp(pT_{\text{delay}})$. We therefore see from (2.8) that Theorem 2.3 applies to our simple model for disease control with TeTrIs (set $\mathbf{H} = \alpha\beta$). However, the true power of Theorem 2.3 is that it holds for any feedback interconnection of the form of (2.7) that satisfies (2.10). This means that the same fundamental limits on performance hold even if we replace our simple model of disease spread from (2.4) with a general compartmental model which predicts an initial period of exponential spread of the disease (if there is no spread, TeTrIs is not really necessary anyway). To see this, suppose that the linearisation of our compartmental model of choice can be written in the general form[5]

$$\frac{dx}{dt} = Ax + BQ, \quad I = Cx. \tag{2.11}$$

[5]This is the general form of the linearisation of a compartmental model

$$\frac{dx}{dt} = f(x,Q), \quad I = g(x).$$

It may seem restrictive that g doesn't depend on Q. However, if it did, this would mean that the effect of quarantining someone would instantly affect whether or not they are infectious, which is rather implausible.

If the model predicts a period of exponential spread of the disease, then the A matrix will have an eigenvalue $p > 0$. Provided this mode is observable and controllable (which would also be necessary for there to be any chance of controlling it through TeTrIs), the transfer function associated with (2.11) will have a pole at p. That is,

$$\hat{I} = \frac{1}{s-p}\mathbf{M}\hat{Q}.$$

Assuming the same model for TeTrIs we can now write the linearisation of the feedback interconnection of (2.5) and (2.11) in the framework of (2.7) by setting $\mathbf{G}_1 = 1/(s-p)\,\mathbf{M}$, and leaving $\mathbf{G}_2 = \alpha \exp(-sT_{\text{delay}})$. The transfer functions in this interconnection also satisfies (2.10), so the same fundamental limit holds. In fact it will continue to hold even if we use more complex models for TeTrIs, provided they still include a total time delay of T_{delay}. We conclude the section with some final remarks on Theorem 2.3.

Remark 2.4. The bound from Theorem 2.3 also applies to the complementary sensitivity function. That is, under the conditions of Theorem 2.3, $\|\mathbf{L}/(1+\mathbf{L})\|_\infty \geq \exp(pT_{\text{delay}})$.

Remark 2.5. Theorem 2.3 continues to hold in the nonlinear setting under the assumption that the feedback interconnection in question has a linearisation. This essentially follows from the fact that the induced $\mathscr{L}_2$-norm of a nonlinear system (the natural generalisation of the H-infinity norm) is always greater than the induced $\mathscr{L}_2$-norm of its linearisation. This effectively shows that by considering the nonlinear effects in more realistic models, performance (as measured using sensitivity functions) can only get worse. This makes it all the more important to aim for performance requirements on the conservative end (i.e. $\|\mathbf{S}\|_\infty \leq 1.2$ rather than $\|\mathbf{S}\|_\infty \leq 2$), necessitating a speedier response. This is discussed in Appendix B.

2.3. Discussion. The purpose of this section has been to expose fundamental limits in epidemic control that arise from the combination of two factors: the natural open-loop instability of the system, and the existence of delays in the feedback loop. Some of our results were stated in general form, but the main motivating example is the stabilization and regulation of an epidemic by means of testing, tracing and isolation of infections. The bounds derived apply to *any* control strategy of this kind, and can be summarized in "the need for speed": if the delays involved in identifying, testing and isolating cases are not very tight, the success of the entire approach is in jeopardy.

There are other strategies for an epidemic control, which are also subject to fundamental limits of this kind. The most commonly deployed one is *social distancing* of the entire population. In the context of the classical SIR models, this means making the parameter β itself a control variable, attempting to stabilize the dynamics at a nonzero number of infections, compatible with the capacity of the healthcare system. Of course, a model of social behavior that would cover the control of β is not easy to obtain, and will not be pursued here. We remark, nonetheless, that for instance a strategy of ordering a lockdown when infections hit a certain threshold is also subject to time delays (due to disease latency times) which will compromise performance.

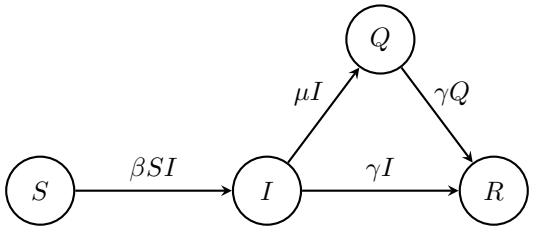
Staying within the realm of contact tracing based control, there is another fundamental limit that will be analyzed in the following section.

3. Track-and-trace: The Need for Scale. The analysis of the preceding section sets the focus on the effect of feedback *delays* in limiting the performance of

the TeTrIs strategy for epidemic control. Here we will address a different limitation of the control strategy that manifests in the presence of disturbances. That is, TeTrIs relies on scarce resources: the availability of technology and trained personnel for taking samples and laboratory testing, for the proactive tracking down of potential infections, and for ensuring appropriate quarantine.

These resources are usually orders of magnitude smaller than the full scale of the population, and thus often saturate in a widespread epidemic such as COVID-19. The question we wish to address is the characterization of these limitations in mathematical models for the epidemic under TeTrIs-based control. To accommodate the nonlinear effect of saturation in a tractable way, we simplify the delay-to-quarantine model to finite dimensional dynamics instead of a pure delay. This alternative is natural in the context of compartmental models: rather than assume that the TeTrIs process takes a fixed amount of time to remove infected people, we assume a rate of removal is given; this can be seen as the macroscopic aggregate of the random times involved in the contract tracing process.

3.1. A Model for Contact Tracing. We thus introduce a compartmental model that incorporates as a *state* the number of people in quarantine Q, in addition to the the standard susceptible (S), infected (I) and removed (R) populations. We assume that people in quarantine effectively isolate and thus are no longer producing new infections.



The TeTrIs control strategy is modeled as follows: Infected people are individually tracked, tested and isolated at a rate μ, meaning that on average, we need a time $1/\mu$ to effectively put these people into quarantine.

Under these assumptions, the dynamics become

$$\frac{d}{dt}\begin{bmatrix} S \\ I \\ Q \\ R \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \\ 0 \\ 0 \end{bmatrix} \beta SI + \begin{bmatrix} 0 \\ -1 \\ 1 \\ 0 \end{bmatrix} \mu I + \begin{bmatrix} 0 \\ -1 \\ 0 \\ 1 \end{bmatrix} \gamma I + \begin{bmatrix} 0 \\ 0 \\ -1 \\ 1 \end{bmatrix} \gamma Q. \tag{3.1}$$

This model was already proposed in [22] and its analysis is simple, since quarantined people can be considered as "early recoveries". More formally, if we consider the dynamics in $\tilde{S} = S, \tilde{I} = I, \tilde{R} = Q + R$, then the model becomes a simple SIR model with recovery rate $\gamma + \mu$ and therefore the critical reproduction rate parameter is

$$R_\mu := \frac{\beta}{\gamma + \mu}. \tag{3.2}$$

In the model without quarantine, the open-loop critical rates is $R_0 = \beta/\gamma$ (corresponding to the case $\mu = 0$). The net effect of contact tracing is to reduce the reproduction rate: $R_\mu < R_0$. In particular, if the contact tracing rate $\mu \to 0$ (contact tracing is extremely slow), it is as if contact tracing is not operating. If contact tracing is extremely fast ($\mu \to \infty$), it can stabilize any open-loop transmission rate.

In fact, the above analysis gives a first rule of thumb to determine the contact tracing speed. That is, provided that the open-loop system is unstable ($R_0 > 1$), we need

$$\frac{1}{\mu} < \frac{1}{\beta - \gamma}, \tag{3.3}$$

i.e., the average isolation time must be controlled. Eq. (3.3) can be compared with (2.6), the main difference stems from the fact that here we are continuously isolating people after a random delay, instead of a fixed one. As an example, if we fix the average recovery time in $1/\gamma = 10$ days and $R_0 = 3$ ($\beta = 0.3$), the average time to isolate is bounded by 5 days.

While this family of quarantining models is well known, we would like to analyze the effect of *saturating* the contact tracing capability. To this end, consider that there is a maximum fraction of the population K that can be tested, tracked, and isolated simultaneously. This can be due to a limit in the total test processing capability, the number of contact tracing agents that are deployed or any combination thereof.

In such a scenario, if the number of infected people is low, then the quarantining rate should be μI, since every infected person is being tracked (equivalently there exists idle tracking and testing capacity). However, if the number of infected people is high ($I > K$), then the quarantining rate should be μK because of the saturation of the control capabilities.

Under these assumptions, the dynamics become

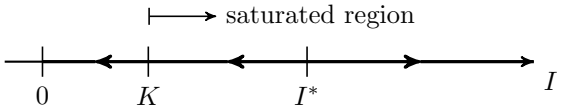
$$\frac{d}{dt}\begin{bmatrix} S \\ I \\ Q \\ R \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \\ 0 \\ 0 \end{bmatrix} \beta SI + \begin{bmatrix} 0 \\ -1 \\ 1 \\ 0 \end{bmatrix} \mu \min\{K, I\} + \begin{bmatrix} 0 \\ -1 \\ 0 \\ 1 \end{bmatrix} \gamma I + \begin{bmatrix} 0 \\ 0 \\ -1 \\ 1 \end{bmatrix} \gamma Q. \tag{3.4}$$

Note that if $K \geq 1$ in (3.4), we recover the first model.

3.2. Understanding the Issue. To highlight the issues introduced by this saturation, we first analyze the dynamics (3.4) under the assumption that $S \approx 1$ (i.e., at the beginning of the epidemic). In that case, the important part of the dynamics is the evolution of infected people, which becomes autonomous:

$$\frac{d}{dt} I = \beta I - \gamma I - \mu \min\{K, I\}. \tag{3.5}$$

The above differential equation is extremely simple to analyze. However, it yields an important insight into the effect of saturation in these kinds of dynamics. Consider the case where $R_0 > 1$, i.e., the system is open-loop unstable, but $R_\mu < 1$, meaning that the system can be stabilized by an "infinite" contact tracing capability, as in (3.1). Then the phase diagram becomes



The new unstable equilibrium that emerges in the approximate dynamics can be readily computed by imposing $dI/dt = 0$ in (3.5) to yield

$$I^* = \frac{\mu K}{\beta - \gamma}. \tag{3.6}$$

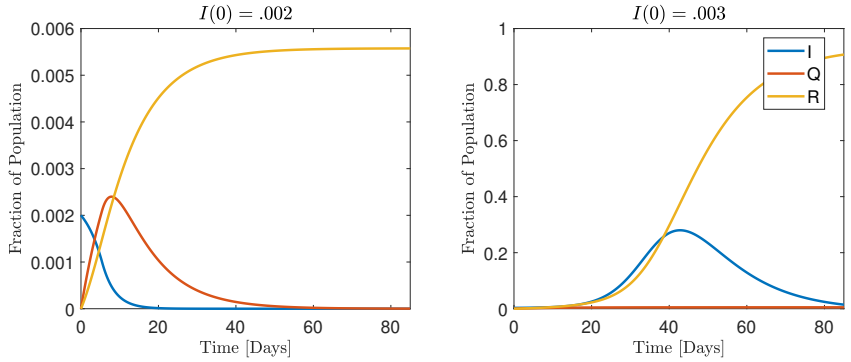


FIG. 5. *Simulation of the system in (3.4) with* $I(0) = 2 \times 10^{-3} < I^*$ *and* $I(0) = 3 \times 10^{-3} > I^*$. *Note the different scales in the y-axis.*

The appearance of this new equilibrium means that the saturation of contact tracing measures leads to a threshold behavior in the number of infected people, a phenomenon already observed in several countries that have lost track of disease spread [15]. Of course, the value I^* is not an equilibrium of the full non-linear dynamics (3.4), but it should operate as a threshold value. We revisit this more formally below.

In addition, using that $R_\mu < 1$, we have $\mu > \beta - \gamma$ and thus $I^* > K$. This means that the stability region is larger than the saturation point of the contact tracing capability. One way to interpret the threshold is to rearrange (3.6) in the following manner:

$$K = \left(\frac{\beta}{\mu} - \frac{\gamma}{\mu}\right) I^*. \tag{3.7}$$

Here the factor $\frac{\beta}{\mu} - \frac{\gamma}{\mu}$ acts as a reproduction number: it can be interpreted as the number of "children" a single infected individual generates until it is traced, minus the ones that recover in that same period. If the total number of new infections generated by a pool I of infected people is larger than the tracing capacity, then the disease will spread in the long run.

Example. To demonstrate the validity of the approximation $S \approx 1$ at the beginning of the epidemic, consider the following scenario: let $\gamma = 1/10$, i.e., recovery time around 10 days and $R_0 = 3$ ($\beta = 0.3$), so the system is open-loop unstable. Assume that we need two days on average to test, trace and isolate people, which amounts to a choice of $\mu = 1/2$. In that case, $I^* = \frac{\mu}{\beta-\gamma} K = 2.5K$, that is, every unit of tracing capability can deal with up to 2.5 simultaneous infections without crossing the threshold. Let us simulate the system for an initial condition with $S \approx 1$. In particular we choose $K = 10^{-3}$, meaning that 1 in 1000 people can be tracked simultaneously. With this choice of K, $I^* = 2.5 \times 10^{-3}$ and we choose $I(0)$ slightly below or above I^*. Results are shown in Fig. 5. We can see that the simulated (nonlinear) system indeed enters the exponential phase immediately after reaching the threshold.

The above analysis, albeit simplistic, illustrates the effects of local non-linearities in the stability behavior of epidemics. Namely, a stable region appears around the extinction equilibrium, but instability can be reinstated if the number of infected people grows large, overwhelming the control capabilities. We now analyze this further

in the complete dynamics (3.4), and then extend the framework to consider the case where the tracing effort is in part spent on contacts that do not become infected.

3.3. Nonlinear Analysis. To understand the effect of the saturation without approximating $S \approx 1$, it is of use to first understand the behavior of $S(t)$. Since, by (3.4), $\frac{d}{dt}S \leq 0$, $S(t)$ is a *decreasing function of time*. This allows us to derive the following monotonicity property for $I(t)$.

PROPOSITION 3.1 (Monotonicity of $I(t)$ under (3.4)). *Consider the dynamics* (3.4). *Then the following property holds:*

$$\frac{d}{dt}I(t_0) < 0 \implies \frac{d}{dt}I(t) < 0, \quad \forall t \geq t_0. \tag{3.8}$$

Proof. Without loss of generality we assume $I(t_0) > 0$. We first consider the case $I(t_0) \leq K$. In this case, it follows from (3.4) that $S(t_0) < 1/R_\mu$. This is the standard scenario where the number of susceptible people is not enough to sustain the epidemic, thus we expect $\frac{d}{dt}I(t) < 0$ for all $t > t_0$.

Indeed, if we assume by contradiction that there is a time t_1 such that $\frac{d}{dt}I(t_1) = 0$ then we get

$$0 = \frac{d}{dt}I(t_1) = (\beta S(t_1) - \gamma - \mu)I(t_1) \implies S(t_1) = \frac{1}{R_\mu} > S(t_0),$$

which contradicts the fact that $S(t)$ is decreasing in time.

The analysis for the case $I(t_0) \geq K$ follows a similar reasoning. Indeed, by considering the saturated version of (3.4), i.e.,

$$\frac{d}{dt}I = \beta SI - \gamma I - \mu K, \tag{3.9}$$

we get that $\frac{d}{dt}I(t_0) < 0$ implies

$$(\beta S(t_0) - \gamma)I(t_0) < \mu K. \tag{3.10}$$

Thus, assuming again by contradiction the existence of t_1, being the first time $\frac{d}{dt}I(t) = 0$ for $t > t_0$, we obtain

$$(\beta S(t_0) - \gamma)I(t_0) < \mu K = (\beta S(t_1) - \gamma)I(t_1) \leq (\beta S(t_0) - \gamma)I(t_1), \tag{3.11}$$

where the first inequality follows from $\frac{d}{dt}I(t_0) < 0$ and the second from the monotonicity of $S(t)$. It follows then that $I(t_1) > I(t_0)$, and therefore

$$0 < I(t_1) - I(t_0) = \int_{t_0}^{t_1} \frac{d}{dt}I(t)dt < 0,$$

where the last inequality holds by the definition of t_1. Thus, such a time t_1 cannot exist. ∎

The preceding proposition illustrates the critical role of the nullcline $\frac{d}{dt}I = 0$ in (3.4) in understanding the threshold behavior in the nonlinear case. To simplify exposition and further understand the role of the nullcline, we consider only the most relevant case when $R_\mu < 1$ and $R_0 > 1$, as before.

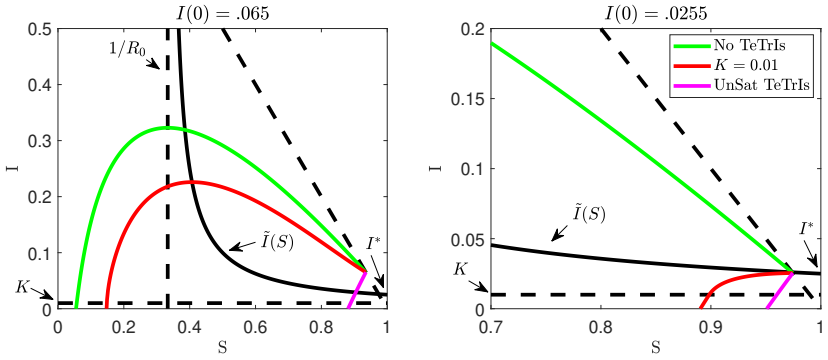


FIG. 6. *S-I region of the phase plane. Trajectories for uncontrolled evolution (green), unsaturated TeTrIs (purple), and TeTrIs with $K = 0.01$ (red) are presented for two initial conditions. On the left, $I(0)$ is above the nullcline and the pandemic spreads. On the right, $I^* < I(0) < \tilde{I}(S(0))$ and the pandemic is contained successfully. The $\tilde{I}(S)$ nullcline (solid black) thus acts as a threshold between successful and unsuccessful TeTrIs.*

In this case, the nullcline is fully within the saturated region, and Proposition 3.1 leads to the simple condition

$$I \le \tilde{I}(S) := \frac{\mu K}{\beta S - \gamma} = \frac{\mu K}{\beta(S - \frac{1}{R_0})} \tag{3.12}$$

for the disease to dissipate without a major outbreak. Indeed, for the number of infectious people to increase, $\frac{d}{dt}I(t)$ must be positive, thus violating (3.12).

A few remarks are in order. First, the threshold is only valid for the range $0 \le \tilde{I}(S) \le 1$. Outside such range, the disease dies out. In particular, $\tilde{I}(S) \ge 0$ leads to the already known $S \le 1/R_0$ condition, and $\tilde{I}(S) \ge 1 \ge I$ guarantees $\frac{d}{dt}I < 0$ for all I. Second, the nonlinear threshold $\tilde{I}(S)$ is a decreasing function of S (see Fig. 6), which implies that the most conservative bound is obtained at $S = 1$, which leads to

$$\tilde{I}(S) = \frac{\mu K}{\beta S - \gamma} \ge \frac{\mu K}{\beta - \gamma} = I^* > K,$$

where the last inequality follows from our assumption $R_\mu < 1$. Thus, the analysis of the previous section leads to a *lower bound* on the critical threshold, which, as expected, is quite accurate when $S \approx 1$.

Example. Consider again the set of parameters $\beta = 0.3$, $\gamma = 1/10$ and $\mu = 1/2$. As mentioned before, since in this case $R_\mu < 1 < R_0$, $\tilde{I}(S) \ge I^* > K$ holds for all S. Fig. 6 considers the case of $K = 0.01$ (red) and compares its trajectory on the (S, I) plane with two additional cases, the unsaturated dynamics ($UnSatTeTrIs$, purple) and the regular dynamics with no track-and-trace ($No\ TeTrIs$, red). On the left, an initial condition $I(0) = 0.65$, $S(0) = 1 - I(0)$, with $I(0)$ above the threshold $\tilde{I}(S)$ (solid black), is considered. On the right, a similar setting but with $I(0) = 0.0255$ between $\tilde{I}(S(0)) = \tilde{I}(0.974) = 0.026$ and $I^* = 0.025$ is considered. This therefore validates the very slight conservativeness in the I^* threshold.

3.4. Modeling the Tracing of Uninfected Contacts. One thing the preceding models do not capture is that the resources of a contact tracing system are also invariably used to test and trace people that have been in contact with infected

individuals, but *have not* developed the infection. As we analyze in this section, the stability region obtained by TeTrIs control policy will be reduced because of this phenomenon.

Consider the following compartmental model for the epidemic spread. As usual, I denotes the infected population at a given time. These infected individuals have multiple contacts which generate secondary infections at rate β, but also have other contacts, say at rate β_1, which do *not* generate infection. Since this classification can only be ascertained by testing, the TeTrIs capability is in part spent on these non-infected contacts. We will denote the population of *potential infections* by P, and separate it from the rest of the susceptible population for which we use the variable S.

For our model, we choose $\beta_1 = \nu\beta$. Here ν can be thought as the "odds ratio" that a contacted individual does not develop the infection. If $\nu = 0$, all potential contacts are infected and the model operates as before, but typically $\nu > 0$, meaning that not all contacts are infected. In particular, in Uruguay where we have access to fine grained data, its value is around $\nu = 10$, meaning that for each infected individual, 10 more people should be tracked.

The open-loop model given below carries out the classification of susceptible individuals into the P and S categories, before incorporating contact tracing:

$$\frac{d}{dt}\begin{bmatrix} S \\ P \\ I \\ R \end{bmatrix} = \begin{bmatrix} -1-\nu \\ \nu \\ 1 \\ 0 \end{bmatrix} \beta IS + \begin{bmatrix} 0 \\ -1 \\ 1 \\ 0 \end{bmatrix} \beta IP + \begin{bmatrix} 0 \\ 0 \\ -1 \\ 1 \end{bmatrix} \gamma I. \tag{3.13}$$

Of course, if we combine both categories of susceptibles into one class $\tilde{S} = S + P$, the model reduces to a classical SIR model with infection rate β and recovery rate γ. Thus the reproduction number for the model in (3.13) is given as before by

$$R_0 = \frac{\beta}{\gamma}.$$

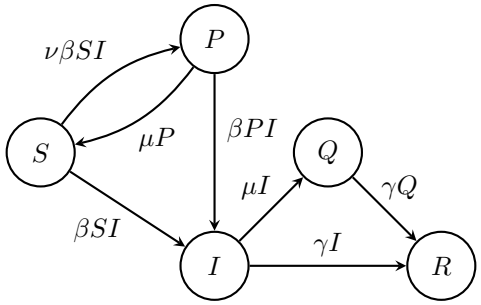
Consider now that the contact tracing effort u is split between u_P and u_I, meaning that the tracking is performed over the whole potentially infected population. Those that are tracked and are infected are isolated, the others are simply "cleared" and return to the normal susceptible class. Adding as before a state variable for quarantined population we obtain the model:

$$\frac{d}{dt}\begin{bmatrix} S \\ P \\ I \\ Q \\ R \end{bmatrix} = \begin{bmatrix} -1-\nu \\ \nu \\ 1 \\ 0 \\ 0 \end{bmatrix} \beta IS + \begin{bmatrix} 0 \\ -1 \\ 1 \\ 0 \\ 0 \end{bmatrix} \beta IP + \begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \\ 0 \end{bmatrix} u_P + \begin{bmatrix} 0 \\ 0 \\ -1 \\ 1 \\ 0 \end{bmatrix} u_I + \begin{bmatrix} 0 \\ 0 \\ -1 \\ 0 \\ 1 \end{bmatrix} \gamma I + \begin{bmatrix} 0 \\ 0 \\ 0 \\ -1 \\ 1 \end{bmatrix} \gamma Q. \tag{3.14}$$

Following the analysis in the previous sections, in the case where there is no limit to the tracing capabilities, we can assume

$$u_P = \mu P, \quad u_I = \mu I, \tag{3.15}$$

where $1/\mu$ is the average time to trace and test one individual, either potential or infected.



Substituting this control law in (3.14), we can easily observe that, since there is no coupling between u_P and u_I, the model reduces to the contact tracing and quarantining model of Section 3.1. Namely, the state $\tilde{S} = S + P, \tilde{I} = I, \tilde{Q} = Q$ and $\tilde{R} = R$ follows exactly the dynamics in (3.1). In particular, the reproduction rate for a given value of μ is the same as in (3.2):

$$R_\mu = \frac{\beta}{\mu + \gamma}. \tag{3.16}$$

Again with sufficiently fast contact tracing, one can cope with any transmission rate.

The interesting case, however, is when contact tracing is limited by the total number of trackers or simultaneous tests that can be performed. Since these tests are performed *before* knowing if a person is a potential infection or an infected individual, the coupling between u_P and u_I becomes

$$u_P + u_I \leqslant \mu K. \tag{3.17}$$

In particular, if we assume that the effort is equally split between all $P + I$ potentially infected individuals, then:

$$u_P(P, I) = \mu \frac{P}{P+I} \min\{P + I, K\} = \mu P \min\left\{1, \frac{K}{P+I}\right\}, \tag{3.18}$$

$$u_I(P, I) = \mu \frac{I}{P+I} \min\{P + I, K\} = \mu I \min\left\{1, \frac{K}{P+I}\right\}. \tag{3.19}$$

Note that $u_P + u_I = \mu \min\{K, P + I\}$ and thus satisfies (3.17). Also when I and P are near zero, the feedback law reduces to (3.15).

3.5. Threshold Analysis. In comparison with (3.4), a full nonlinear analysis in this case is more involved. Therefore, we resort to the strategy of analyzing the behavior of the saturated policy around the disease free equilibrium where $S \approx 1$. In this setting, $P \ll 1$ and $I \ll 1$ so the product term IP can be disregarded.[6] Substituting this condition and the control law (3.18) in (3.14), the dynamics become autonomous in P and I with

$$\frac{d}{dt}\begin{bmatrix} P \\ I \end{bmatrix} = \begin{bmatrix} 0 & \nu\beta \\ 0 & \beta - \gamma \end{bmatrix} \begin{bmatrix} P \\ I \end{bmatrix} - \mu \min\left\{1, \frac{K}{P+I}\right\} \begin{bmatrix} P \\ I \end{bmatrix}. \tag{3.20}$$

We have the following:

[6] This is equivalent to considering that every potential contact only arises from a single infected interaction.

PROPOSITION 3.2. *Under the condition $R_0 > 1$ (uncontrolled open loop) and $R_\mu < 1$, the dynamics in (3.20) have a locally asymptotically stable disease free equilibrium $P = I = 0$, and a further unstable equilibrium emerges at*

$$P^* = \frac{\nu\beta}{((1+\nu)\beta-\gamma)(\beta-\gamma)}\mu K, \quad I^* = \frac{1}{(1+\nu)\beta-\gamma}\mu K. \tag{3.21}$$

Proof. We begin by analyzing the disease free case, which is readily verified to be an equilibrium after substitution in (3.20). The Jacobian matrix in this case retains a diagonal term $-\mu$ since the saturation is not in effect near the origin. Thus the Jacobian is

$$J_1 = \begin{bmatrix} -\mu & \nu\beta \\ 0 & \beta-\gamma-\mu \end{bmatrix}.$$

The Jacobian has two eigenvalues, $-\mu < 0$ and $\beta-\gamma-\mu$ that is also negative because of the assumption that $R_\mu < 1$, hence the equilibrium is locally stable.

To find the second equilibrium, we assume that the saturation is active and imposes equilibrium in (3.20):

$$\begin{bmatrix} 0 & \nu\beta \\ 0 & \beta-\gamma \end{bmatrix} \begin{bmatrix} P^* \\ I^* \end{bmatrix} - \mu \frac{K}{P^*+I^*} \begin{bmatrix} P^* \\ I^* \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

After some algebra one arrives at the expressions in (3.21) for P^* and I^*.

Furthermore,

$$P^* + I^* = \frac{\mu}{\beta-\gamma} K > K, \tag{3.22}$$

under the hypothesis that $\mu > \nu\beta - \gamma \Leftrightarrow R_\mu < 1$. Hence, for any testing rate that stabilizes under infinite contact tracing assumptions, one gets an unstable equilibrium when the saturation comes into play. Moreover, note that the total number being tracked at this new equilibrium coincides with the threshold (3.6).

That this equilibrium is indeed unstable can be seen by analyzing its Jacobian matrix

$$J_2 = \begin{bmatrix} 0 & \nu\beta \\ 0 & \beta-\gamma \end{bmatrix},$$

which corresponds to the open-loop model that has a positive eigenvalue $\beta - \gamma > 0$ under the assumption $R_0 > 1$. ∎

As a final remark, note that the equilibrium (3.21) verifies

$$\frac{P^*}{I^*} = \frac{\nu\beta}{\beta-\gamma} = \frac{R_0}{R_0-1}\nu. \tag{3.23}$$

This supports the intuitive observation that, when ν is large, most of the contact tracing effort is spent only on the potential contacts, reducing the stability margin. Below we analyze this in a numerical example.

Example. To depict the behavior of the dynamics (3.20), we choose as before $\gamma = 1/10$ (10 days average recovery time) and $\beta = 3\gamma$, yielding $R_0 = 3$. The ratio ν is taken as $\nu = 10$ as observed in some cases, consistent with current measurements

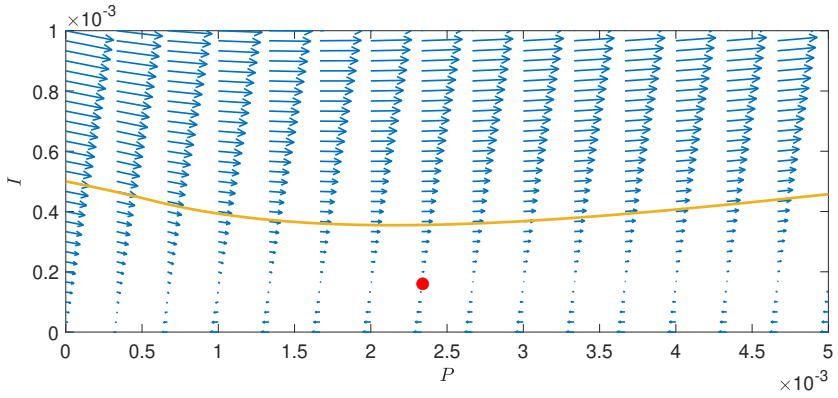


FIG. 7. *Phase diagram of* (3.20) *and unstable equilibrium point of the approximate dynamics. We superimpose the solution of the nonlinear version depicted in Fig. 8.*

in the real epidemiological scenario in Uruguay, where approximately 10 contacts are traced per infected individual, generating only one new infection.

If we assume that $K = 10^{-3}$, meaning that 1 in 1000 people can be tracked and tested simultaneously, then the unstable equilibrium occurs at

$$P^* + I^* = 2.5 \times 10^{-3},$$

but with a lower number of infections, namely,

$$P^* = 2.34 \times 10^{-3}, \quad I^* = 0.16 \times 10^{-3}.$$

Observe that these parameters are also consistent with the numerical example in Section 3.1, where the stability threshold was at $I = 2.5 \times 10^{-3}$. Now that the contact tracing is burdened with potential contacts, and the stability region diminishes in consequence.

The phase plot is depicted in Fig. 7. In particular, starting from an initial condition $I(0) = 0.5 \times 10^{-3}$ (which would be clearly stable in (3.4)) and $P(0) = 0$, the system enters the exponential phase due to the secondary contacts that burden the contact tracing capabilities. In particular, in Fig. 8 we can observe that at the peak 70% of the population becomes a potential contact simultaneously, and the susceptible people go quickly to 0, meaning that the whole population has been in contact with an infected individual, clearly overwhelming the tracking and testing capabilities.

3.6. Discussion. To conclude this section, let us recap the main results derived. The first result is that, whenever there is a cap on the contact tracing capability, a threshold behavior develops in the dynamics. This emphasizes the *need for scale*, summarized succinctly in (3.6) and its nonlinear counterpart (3.12). Whenever the infected number grows, the testing and tracing capacity should grow linearly with the number of infections in order to avoid saturation. On the other hand, the system can work in the saturated regime without becoming overwhelmed, but once the threshold is crossed the epidemic will spread.

The second result is that this stability margin is greatly compromised by the fact that testing and tracing capacity is burdened with the need of following contacts that do not become infected. This is summarized in (3.22) and (3.23), that evidence how

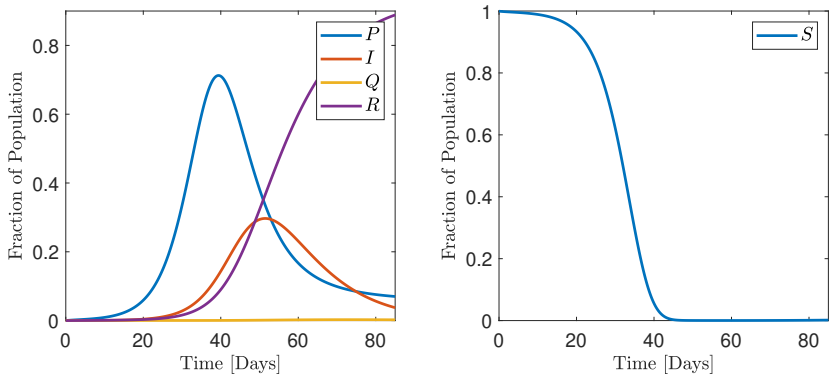


FIG. 8. *Unstable trajectories of the saturated system with limited contact tracing.*

saturation comes into play due to the total number of contacts, and that this total number is dominated by potential contacts.

4. Conclusions. This work presents a cautionary message of the fundamental limits involved in preventing disease propagation during an epidemic. Our results highlight the particularly dangerous combination of instability and non-linearity, intrinsic of the disease spread process (our plant), together with delays and capacity constraints, intrinsic of the TeTrIs process (our actuator), that makes the disease control problem fundamentally challenging. It is important to notice that some of our quantitative predictions are, to a certain extent, pessimistic, as we only consider one method for disease spread prevention, i.e., TeTrIs. Clearly, complementing such a process with other control mechanisms, such as social distancing, using masks, etc., can improve the effectiveness and robustness of the disease spread mitigation efforts. Nevertheless, irrespective of the methods used, we believe that the needs for speed and scale are, at its core, necessary for effective disease prevention.

REFERENCES

[1] *Contact tracing in the context of covid-19*, May 2020, https://www.who.int/publications/i/item/contact-tracing-in-the-context-of-covid-19. Accessed: 2020-10-30.
[2] K. J. ÅSTRÖM, *Limitations on control system performance*, European Journal of Control, 6 (2000), pp. 2–20.
[3] K. J. ÅSTRÖM AND R. M. MURRAY, *Feedback systems: an introduction for scientists and engineers*, Princeton university press, 2 ed., 2010.
[4] F. CASELLA, *Can the covid-19 epidemic be controlled on the basis of daily test reports?*, IEEE Control Systems Letters, 5 (2020), pp. 1079–1084.
[5] M. CHAN-YEUNG AND R.-H. XU, *Sars: epidemiology*, Respirology, 8 (2003), pp. S9–S14.
[6] J. CLARKE, *Contact tracing for chlamydia: data on effectiveness*, International journal of STD & AIDS, 9 (1998), pp. 187–191.
[7] E. DONG, H. DU, AND L. GARDNER, *An interactive web-based dashboard to track covid-19 in real time*, The Lancet infectious diseases, 20 (2020), pp. 533–534.
[8] J. C. DOYLE, B. A. FRANCIS, AND A. R. TANNENBAUM, *Feedback control theory*, Macmillan Publishing Co, 1992.
[9] K. T. EAMES AND M. J. KEELING, *Contact tracing and disease control*, Proceedings of the Royal Society of London. Series B: Biological Sciences, 270 (2003), pp. 2565–2571.
[10] N. FERGUSON, D. LAYDON, G. NEDJATI GILANI, N. IMAI, K. AINSLIE, M. BAGUELIN, S. BHATIA, A. BOONYASIRI, Z. CUCUNUBA PEREZ, G. CUOMO-DANNENBURG, ET AL., *Report 9: Impact of non-pharmaceutical interventions (npis) to reduce covid19 mortality and healthcare de-*

mand, (2020).
[11] N. M. Ferguson, D. A. Cummings, C. Fraser, J. C. Cajka, P. C. Cooley, and D. S. Burke, *Strategies for mitigating an influenza pandemic*, Nature, 442 (2006), pp. 448–452.
[12] M. FitzGerald, D. Thirlby, and C. Bedford, *The outcome of contact tracing for gonorrhoea in the united kingdom*, International journal of STD & AIDS, 9 (1998), pp. 657–660.
[13] C. Foias, A. Tannenbaum, and G. Zames, *Weighted sensitivity minimization for delay systems*, IEEE Transactions on Automatic Control, 31 (1986), pp. 763–766.
[14] C. Fraser, S. Riley, R. M. Anderson, and N. M. Ferguson, *Factors that make an infectious disease outbreak controllable*, Proceedings of the National Academy of Sciences, 101 (2004), pp. 6146–6151.
[15] J. Hasell, E. Mathieu, D. Beltekian, B. Macdonald, C. Giattino, E. Ortiz-Ospina, M. Roser, and H. Ritchie, *A cross-country database of covid-19 testing*, Scientific data, 7 (2020), pp. 1–7.
[16] W. O. Kermack and A. G. McKendrick, *A contribution to the mathematical theory of epidemics*, Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character, 115 (1927), pp. 700–721.
[17] P. Khargonekar and A. Tannenbaum, *Non-euclidian metrics and the robust stabilization of systems with parameter uncertainty*, IEEE Transactions on Automatic Control, 30 (1985), pp. 1005–1013, https://doi.org/10.1109/TAC.1985.1103805.
[18] I. Z. Kiss, D. M. Green, and R. R. Kao, *The effect of network mixing patterns on epidemic dynamics and the efficacy of disease contact tracing*, Journal of the Royal Society Interface, 5 (2008), pp. 791–799.
[19] B. A. Macke and J. E. Maher, *Partner notification in the united states: an evidence-based review*, American Journal of Preventive Medicine, 17 (1999), pp. 230–242.
[20] J. Müller and B. Koopmann, *The effect of delay on contact tracing*, Mathematical biosciences, 282 (2016), pp. 204–214.
[21] J. Müller, M. Kretzschmar, and K. Dietz, *Contact tracing in stochastic and deterministic epidemic models*, Mathematical biosciences, 164 (2000), pp. 39–64.
[22] M. Nuno, C. Castillo-Chavez, Z. Feng, and M. Martcheva, *Mathematical models of influenza: the role of cross-immunity, quarantine and age-structure*, in Mathematical Epidemiology, Springer, 2008, pp. 349–364.
[23] Y. J. Park, Y. J. Choe, O. Park, S. Y. Park, Y.-M. Kim, J. Kim, S. Kweon, Y. Woo, J. Gwack, S. S. Kim, et al., *Contact tracing during coronavirus disease outbreak, south korea, 2020*, Emerging infectious diseases, 26 (2020), pp. 2465–2468.
[24] G. Stein, *Respect the unstable*, IEEE Control Systems Magazine, 23 (2003), pp. 12–25.

Appendix A. Proof of Theorem 2.1.

We begin by linearising the model in (2.4) and (2.5). Eliminating S using the algebraic equation in (2.4) and then linearising about the point $(I, R, Q) = (0, 0, 0)$ shows that for small deviations,

$$\frac{dI}{dt} = (\beta - \gamma)\, I - \beta Q. \tag{A.1}$$

Equation (2.5) is already linear. We are therefore required to show that the interconnection of (A.1) and (2.5) is stable. Eliminating Q from the I equation in (A.1) with (2.5) gives

$$\frac{dI}{dt} + \gamma I + \alpha\beta \exp\left(-\gamma T_{\text{delay}}\right) I\left(t - T_{\text{delay}}\right) - \beta I = 0.$$

Stability is then equivalent to all the roots of the characteristic equation lying in the open left-half-plane. That is,

$$s + \gamma + \alpha\beta \exp\left(-\gamma T_{\text{delay}}\right) \exp\left(-s T_{\text{delay}}\right) - \beta \neq 0,\ \forall s \in \overline{\mathbb{C}}_+.$$

Putting $\tilde{s} = s/\beta$ and rearranging shows that this is equivalent to

$$\tilde{s} + R_0^{-1} + \alpha \exp\left(-\beta T_{\text{delay}}\left(\tilde{s} + R_0^{-1}\right)\right) \neq 1,\ \forall \tilde{s} \in \overline{\mathbb{C}}_+. \tag{A.2}$$

A standard Nyquist argument then shows that this holds if and only if the curve given by

$$f\left(\tilde{s}\right) \coloneqq \tilde{s} + R_0^{-1} + \alpha \exp\left(-\beta T_{\text{delay}}\left(\tilde{s} + R_0^{-1}\right)\right)$$

when evaluated along the usual Nyquist D-contour does not encircle 1. A simple sufficient condition for this is that

(i) $f(0) > 1$;

(ii) $\frac{d}{d\omega}\left(\mathrm{Im}\left(f\left(j\omega\right)\right)\right) > 0$;

since together (i)–(ii) ensure that the curve only crosses the real axis to the right of 1 (technically we also need to consider the real axis crossing on the return arc along the D-contour, but since for large s, $f(s) \approx s$, these will be to the right of 1). It is readily checked that (i) is equivalent to the condition from the theorem statement. That is,

$$\text{(i)} \quad \iff \quad T_{\mathrm{delay}} < \frac{1}{\gamma} \ln\left(\frac{\alpha}{1 - R_0^{-1}}\right) =: T^*.$$

For (ii), observe that

$$\frac{d}{d\omega}\left(\mathrm{Im}\left(f\left(j\omega\right)\right)\right) = 1 - \alpha\beta T_{\mathrm{delay}} \exp\left(-\beta T_{\mathrm{delay}} R_0^{-1}\right) \cos\left(\beta T_{\mathrm{delay}} \omega\right).$$

Therefore, it is sufficient that $\alpha\beta T_{\mathrm{delay}} \exp\left(-\beta T_{\mathrm{delay}} R_0^{-1}\right) < 1$. We will demonstrate this in two stages. First observe that $\alpha\beta T_{\mathrm{delay}} \exp\left(-\beta T_{\mathrm{delay}} R_0^{-1}\right) \leq \alpha R_0 \exp\left(-1\right)$. Therefore, if $R_0 < \exp(1)$, (ii) holds (recall that $0 \leq \alpha \leq 1$). Now assume that $R_0 \geq \exp(1)$. We then see that if this is the case,

$$\ln\left(\frac{\alpha}{1 - R_0^{-1}}\right) \leq \ln\left(\frac{1}{1 - \exp\left(-1\right)}\right) \approx 0.5 < 1,$$

so (i) implies that

$$\beta T_{\mathrm{delay}} < \beta T^* < R_0.$$

Next observe that for $x < R_0$, the function $x \exp\left(-x/R_0\right)$ is monotonically increasing in x. Therefore,

$$\begin{aligned} \alpha\beta T_{\mathrm{delay}} \exp\left(-\beta T_{\mathrm{delay}} R_0^{-1}\right) &< \alpha\beta T^* \exp\left(-\beta T^* R_0^{-1}\right) \\ &= R_0\left(1 - R_0^{-1}\right) \ln\left(\frac{\alpha}{1 - R_0^{-1}}\right) \leq 1. \end{aligned}$$

Therefore, (i) $\implies$ (ii), and by consequence the conditions of the theorem are sufficient for stability. Necessity follows since if $T_{\mathrm{delay}} \geq T^*$, then $f(0) \leq 1$. Since for $x \gg 0$, $f(x) > 1$, by the intermediate value theorem there must be some $x \geq 0$ for which $f(x) = 1$. Therefore, Equation (A.2) does not hold, and the system will be unstable.

Appendix B. Extending Theorem 2.3 to the Nonlinear Setting.

In this section we will demonstrate that under appropriate assumptions, a natural analogue of Theorem 2.3 holds in the nonlinear setting. To do this we will prove that the induced $\mathscr{L}_2$-norm of a system is always lower-bounded by the induced $\mathscr{L}_2$-norm of its linearisation. Since the induced $\mathscr{L}_2$-norm of an LTI system is equal to its H-infinity norm, this shows that if the linearisation of a nonlinear system is LTI, then the induced $\mathscr{L}_2$-norm of the sensitivity function of the nonlinear system must satisfy the same bound from Theorem 2.3.

The result we are trying to prove is in fact rather elementary. However, it requires a bit of setup to lay out the appropriate definitions and concepts. The difficulties stem from the fact that we would like to combine nonlinear state-space models (to describe general compartmental models for disease spread) and delays. Accordingly we adopt

the standard operator theoretic setup on $\mathscr{L}_2$ which covers both these types of model. More specifically, $\mathscr{L}_2$ is the space of functions $f : [0, \infty) \to \mathbb{R}$ with finite norm

$$\|f\| \coloneqq \sqrt{\int_0^\infty |f(t)|^2 \, dt}.$$

This is a subspace of $\mathscr{L}_{2e}$, whose members need only be square integrable on finite intervals. An operator is a function $\mathcal{G} : \mathscr{L}_{2e} \to \mathscr{L}_{2e}$, and the induced $\mathscr{L}_2$-norm of an operator is defined as

$$\|\mathcal{G}\|_{\mathscr{L}_2} \coloneqq \sup \left\{ \frac{\|\mathcal{G}(u)\|}{\|u\|} : u \in \mathscr{L}_{2e}, u \neq 0 \right\}.$$

In the case where the operator $\mathcal{G}$ is describing the dynamics of a LTI system with transfer function $\mathbf{G}$, $\|\mathcal{G}\|_{\mathscr{L}_2} = \|\mathbf{G}\|_\infty$.

The natural generalisation of a linearisation in this setting is given by the Fréchet derivative. An operator $\mathcal{G}$ is Fréchet differentiable at a point $x \in \mathscr{L}_2$ if there exists a linear operator $\mathcal{A}$ such that

$$\lim_{h \to 0} \frac{\|\mathcal{G}(x+h) - \mathcal{G}(x) - \mathcal{A}(h)\|}{\|h\|} = 0.$$

If such a linear operator exists, it is unique, and we denote the Fréchet derivative of $\mathcal{G}$ at x as $D\mathcal{G}(x) = \mathcal{A}$.

With these definitions in place, we are ready to state the main result of this section. The following lemma shows that provided the linearisation exists, the induced $\mathscr{L}_2$-norm of the linearisation of an operator about a fixed point (an equilibrium point) is always smaller than the $\mathscr{L}_2$-norm of the operator itself. This means that if we have a nonlinear system $\mathcal{G}$ with linearisation described by an LTI system with transfer function $\mathbf{G}$, then $\|\mathcal{G}\|_{\mathscr{L}_2} \geq \|\mathbf{G}\|_\infty$. This immediately gives us a nonlinear generalisation of Theorem 2.3. In particular, if we instead study the nonlinear feedback interconnection

$$\begin{aligned} e_i &= \mathcal{G}_i(e_{i-1}) + d_i, \ i \in \{1, \dots, n\} \\ e_0 &= -e_n, \end{aligned} \tag{B.1}$$

and define the sensitivity functions to be the operators $\mathcal{S}_i : d_i \to e_i$, then provided the linearisations of $\mathcal{S}_i$ are LTI, $\|\mathcal{S}_i\|_{\mathscr{L}_2}$ must satisfy exactly the same lower bound from Theorem 2.3.

LEMMA B.1. *Given an operator $\mathcal{G}$, if $\mathcal{G}(0) = 0$ and $\mathcal{G}$ is Fréchet differentiable at 0, then*

$$\|\mathcal{G}\|_{\mathscr{L}_2} \geq \|D\mathcal{G}(0)\|_{\mathscr{L}_2}.$$

Proof. Let $\mathcal{A} = D\mathcal{G}(0)$. Using the reverse triangle inequality shows that for any non-zero $x \in \mathscr{L}_{2e}$ and non-zero $\epsilon \in \mathbb{R}$,

$$\begin{aligned} \|\mathcal{G}\|_{\mathscr{L}_2} \geq \|\mathcal{G}(\epsilon x)\| / \|\epsilon x\| &= \|\mathcal{G}(\epsilon x) - \mathcal{A}(\epsilon x) + \mathcal{A}(\epsilon x)\| / \|\epsilon x\| \\ &\geq \|\mathcal{A}(x)\| / \|x\| - \|\mathcal{G}(\epsilon x) - \mathcal{A}(\epsilon x)\| / \|\epsilon x\|. \end{aligned}$$

Taking the limit $\epsilon \to 0$, we see from the definition of the Fréchet derivative that this implies $\|\mathcal{G}\|_{\mathscr{L}_2} \geq \|\mathcal{A}(x)\| / \|x\|$. Taking the sup over $x \in \mathscr{L}_{2e}$ gives the result. ∎