

# Sustainable Commons Production and the Virtues of Moderation

James Grimmelmann  
TPRC, September 2007

INTRODUCTION.....	1
I. BACKGROUND.....	1
II. MODERATION.....	8
III. CASE STUDIES.....	22
CONCLUSION.....	30

## Introduction

Can online communities remain free and open for long, or will they instead inevitably retreat into closure and control? Economic theories of the commons seemingly support both sides; one can make a reasonable case both that widespread freedom to participate in an online commons encourages massive cooperation, and that it encourages massive abuse. Both claims are correct; both phenomena real. Online communities are semicommons; rival network infrastructure is used to exchange nonrival information goods. By looking more closely at how actual communities balance these pressures, we can see that neither freedom nor control are absolutes, that moderators have a robust toolkit of techniques to encourage healthy collaboration, and that well-moderated communities can preserve substantial freedom for their participants even in the face of threatened abuse.

## I. Background

### *Public and Private Goods*

Economic analysis of the commons begins with the distinction between private goods, which are rival and excludable, and pure public goods, which are neither. On both the rival and non-rival side of the ledger, traditional analysis would generally have counseled promoting excludability. Rival but nonexcludable goods are subject to wasteful racing behavior as different potential users seek to make use of them before someone else does. The result is the so-called “tragedy of the commons,” as the resource is depleted through competitive overuse. This depletion is fundamentally a problem of nonexcludability; where competing users can be kept from the good, the race can be terminated by whoever has the power to exclude. The traditional response thus to find some way of regenerating excludability, such as private property rights or government regulation.

On the other hand, goods that are nonrival and also nonexcludable generate a problem of undersupply, since there is little possibility of a market in such goods. As soon as more than one person has the good, competition between them will drive the price to zero. The traditional response is to invest in making the good more excludable, in the hopes of creating a toll good for which the production incentives are at least present, even if inefficiently sized (given the impossibility of setting price equal to marginal cost or of perfectly price discriminating). This is one explanation of copyright, for example, which creates a form of legal excludability. The exclusion is typically costly, and so the costs of exclusion must be balanced against the incentive gains for production.

To sum up, pure and impure public goods can present two distinctive problems. First, there is the problem of waste, which is most acute when the good is rival but nonexcludable. Second, there is the problem of insufficient incentives to invest, which arises when the good is nonrival and was traditionally considered most acute when the good was also nonexcludable. Thus, prior to the modern flourishing of commons theory, one might naturally have believed that the best response to both problems was to focus on creating excludability. In one context, commons theory has embraced that response; in another, commons theory has repudiated it.

### ***Tragedy***

In 1968, Garrett Hardin's influential article "The Tragedy of the Commons" appeared in *Science* with essentially the argument given above about wasteful overuse of common-pool resources.<sup>1</sup> His specific subject was overpopulation, but he presented examples from parking meters to pollution and delivered the sobering verdict that:

Adding together the component partial utilities, the rational herdsman concludes that the only sensible course for him to pursue is to add another animal to his herd. And another; and another. . . . But this is the conclusion reached by each and every rational herdsman sharing a commons. Therein is the tragedy. Each man is locked into a system that compels him to increase his herd without limit--in a world that is limited. Ruin is the destination toward which all men rush, each pursuing his own best interest in a society that believes in the freedom of the commons. Freedom in a commons brings ruin to all.<sup>2</sup>

---

<sup>1</sup> Garrett Hardin, *The Tragedy of the Commons*, 162 *SCIENCE* 1243 (1968).

<sup>2</sup> *Id.* at 1244.

Against this tragic fate, Hardin saw only one effective response: coercion, either in the form of “private property” in the formerly common resource or “allocat[ion] of the right to enter.”<sup>3</sup>

Subsequent research on common-pool resources has found that sufficiently coercive control over resource usage is compatible with common ownership. Many natural and man-made resources are in fact held communally (whether by law or in practice) without suffering this supposedly inexorably tragic fate. The modern synthesis view is that such cases involve communities that have succeeded in creating and then enforcing on themselves a regime of limited access to the resource. Unlike property rights or direct regulation, both of which are institutions supplied by the state, these community management regimes are bottom-up and are at least initially the product of voluntary cooperation. Nonetheless, they do effectively supply Hardin’s “coercion;” community members who violate the community’s access rules will be punished by they community. We can call this revised explanation the Tragic story because it remains within the basic parameters of Hardin’s explanation of the Tragedy of the Commons.

The Tragic story emphasizes that not all circumstances are right for common ownership. Exact lists differ, but there is by now a rough consensus on the core factors that distinguish successful from unsuccessful common ownership regimes. The community needs both good institutions to gather information about the resource and its usage and a forum to discuss its management. There should be graduated sanctions available, starting with small shaming penalties for minor or first infractions against the access rules and escalating to severe penalties for sustained and flagrant overuse. The community must be able to participate in making and enforcing the rules, at multiple levels of generality. And, perhaps most critically, the community itself should be well-bounded, so that the set of potential users is known, mostly closed to new entrants, and ideally small. While large common-ownership groups do sometimes exist, they typically involve the aggregation of smaller well-functioning units. The small, close-knit group is the model for common-pool resource theory; large and diffuse groups are generally seen as poor candidates for common ownership. The expectation is that they will fail to solve collective action problems and will fall back into wasteful overuse. Remarkably enough, another strain of commons theory has come to almost exactly the opposite conclusion.

---

<sup>3</sup> Id. at 1245.

## *Comedy*

The conventional economic apology for intellectual property, described above, has long coexisted with a strong critique: once an (idealized) intellectual good exists, nonrivalry means that it can be shared with the world at no cost, and nonexclusivity means that it will be. Thus, the legal exclusivity at the heart of intellectual property law is an artificial state-created scarcity that pits *ex ante* incentives to create against the *ex post* value of broad access. There is, however, a deeper critique. Intellectual goods are not simply consumed; they are also essential inputs into the production of other intellectual goods. Overprotection poses a critical additional danger to future creativity, and lower protection can increase creativity by speeding the circulation of ideas.<sup>4</sup> The public domain becomes not simply a negative space of unprotected works and inventions, but a positive resource of immense richness that is and should be held open to all: a commons. Indeed, because there is no problem of rival consumption and thus no danger of waste, it becomes a perfectly functioning commons.

Carol Rose's remarkable *The Comedy of the Commons*<sup>5</sup> combines this point with an observation about social uses of resources. By looking at the historical law of "inherently public property," particularly roads and waterways, Rose developed a theory of the role of the public (distinct from government) as an owner and manager of property. She pointed to the positive network effects scale returns associated with increased use of certain public spaces. A dance in a public square becomes more enjoyable for all participants as it becomes more popular; commerce along a road network becomes more fruitful for all traders as the number of trading partners increases. She then combined this observation with the holdout point to create a positive theory of the benefits of common ownership:

Suppose that a private individual owned a traditional festival ground and that, at least for the festival day, the local residents placed a higher value on this festival use than could be taken from any alternative uses. Ownership of this unique property would give the owner a classic opportunity for rent capture.

But what created the "rent"? The very "publicness" of the festival use; its non-exclusivity makes it valuable, because this activity is exponentially enhanced by greater participation. This value is what customary doctrines refused to permit

---

<sup>4</sup> This account sometimes conceptualizes the problem created by overprotection as an *anticommons*, a space in which many actors have the ability to prevent action, so that nothing is ever done because the transaction costs of negotiating all of the necessary permissions would be prohibitive.

<sup>5</sup> Carol Rose, *The Comedy of the Commons: Custom, Commerce, and Inherently Public Property*, 53 U. CHI. L. REV. 711 (1986).

a private owner to tap or to thwart. In fact, the usual rationing function of pricing would be counterproductive here: participants need encouragement to join these activities, where their participation produces beneficial “externalities” for other participants.”<sup>6</sup>

This is an important inversion. It is not just a theory about the dangers of holdouts; it is also a theory of *self-generated incentives for investment*. If the holdout problem can be solved and the costs of participation driven down, the value created for each individual by contact with others becomes sufficient incentive for her to participate, and so on in a virtuous cycle.

This point has resonated powerfully with theorists of the intellectual commons. Writers such as Yochai Benkler, Rishab Ghosh, and Steve Weber have applied this mode of analysis to open source software and other “impossible public goods.” They have argued that voluntary contribution to an intellectual commons can indeed be self-sustaining and that Rose-style scale returns are one important factor in overcoming the seemingly intractable incentive problem. Social, signaling, and intrinsic incentives may not necessarily be large for many participants, but they are large enough to elicit participation, which is all that matters. Common access lowers costs beneath the threshold at which broad participation becomes feasible, and thereby harnesses scale returns from increased participation. In many domains, all that is required is to throw the gates open, to make the common community of creators and users as large as possible.

### ***Layering***

Thus, one strain of commons theory says to restrict access to a small community; another says to make the community enormous. There is no direct conflict: the first strain by its terms applies to rival resources and the second to nonrival ones. This is all well and good, but the most pressing problems for those who care about such things do not involve resources that can be easily placed into one bin or the other. Indeed, the pure theory of the intellectual commons only directly applies as such to information itself. Given that information must always be instantiated to be used or communicated—whether noted in a human mind with limited attention, stored on a tangible object, or transmitted through an online channel with limited bandwidth—the Tragic story is always close at hand, ready to apply to these rival instantiations. At the same time, it seems clear that something more is going on than simple competition to exhaust the rival

---

6 Id. at 768–69.

components of instantiated information. Scholars have developed one further idea that preserves the possibility that both stories could apply simultaneously: layering.

The term comes from computer science, where it applies to the different “layers” involved in a computer network. Two computers may use SMTP to exchange email over a network that runs TCP/IP over 1000-BASE-T gigabit Ethernet over Category 5e twisted pair copper cables, but these physical and lower-level networking details are irrelevant from the point of view of the email program that the owners of these computers use to exchange messages. The wires (the “physical” layer), the Ethernet protocol (the “link” layer), the TCP/IP protocol (the “transport” layer), the SMTP email protocol (the “application” layer), and the contents of the emails themselves (the “content” layer) are distinct. Email programs can work perfectly well despite having no knowledge of how exponential backoff works, or even that the network, multiple layers down, uses Ethernet.

Layering permits different resource allocation regimes at different layers. In particular, one can make a system controlled at one layer but free at a higher layer. The same network may be fully private at the physical layer (only the company IT manager can enter the room with the server), a limited-access common-pool-resource at the link and transport layers (only employees in the building can connect to it, but they can do so freely), an open-to-the-world common-pool-resource at the application layer (the company provides free ad-supported image hosting to Internet users at large), and something approaching a true commons at the content layer (the company exercises no control over what images users create and share). This form of layered sharing is visible in many communications systems, but is a particularly prevalent feature of the Internet. Indeed, a number of scholars give it credit for the Internet’s remarkable success. They see an application layer largely open to new innovations and a content layer largely open to all forms of communication and sharing, both shining examples of the Comedic story.

Of course, one cannot manufacture something from nothing and exceed one layer’s capacity simply by adding more layers atop it. Where capacity is a serious problem, the Tragic story is also plausible. Simplifying greatly, one might say that many debates over Internet law and policy have been driven by a debate between those who tell a Comedic story about the higher layers and those who tell a Tragic story about the lower ones. Comedians consider the productive freedoms at the higher layers so valuable that they require legal protections, with waste problems at the lower layers either practically irrelevant due to surplus capacity or a proper subject for

specific, targeted responses that leave higher-level freedoms intact. Tragedians instead see the lower-level waste issues as fundamental, so that providers will be unwilling to invest in creating capacity unless they have sufficient legal protection. This dynamic is clearly visible in the debates over spectrum allocation, network neutrality, and trespass to computer systems. Comedians fear that infrastructure will try to capture the gains from the higher-level commons, thereby killing the goose that lays the golden eggs; Tragedians fear that if there is too much solicitousness of the higher-level commons there will be no infrastructure.

### ***Semicommons***

To make further progress in these debates and related ones, we need to focus more closely on the interactions between layers—in particular, on those mechanisms that moderate use of the higher layers. Almost every communications system, even the most seemingly unrestrained, has some such mechanisms in place, ranging from highly informal social norms to quite detailed and mechanistic pricing schemes. These moderation strategies are best understood as techniques to prevent Tragic overuse of the higher layers in ways that would tax the capacity of lower layers without abandoning the Comedic virtues at higher layers:

Henry Smith’s concept of a semicommons provides the natural abstraction:

In a semicommons, a resource is owned and used in common for one major purpose, but, with respect to some other major purpose, individual economic units—individuals, families, or firms—have property rights to separate pieces of the commons. Most property mixes elements of common and private ownership, but one or the other dominates. . . . In what I am calling a semicommons, both common and private uses are important and impact significantly on each other.<sup>7</sup>

Smith’s “archetypal example of a semicommons” is the common-pasturage system of medieval Europe. Sheep could be grazed freely across the fields of a village during fallow seasons, but during growing seasons, individual farmers had exclusive rights to their strips of land.

This framework directs our attention to several important factors. First, it teaches that the externalities in a semicommons come in distinct flavors, depending on whether the source of the effect is a common or a private attribute, and whether the target of the effect is a common or a private attribute.<sup>8</sup> Third, it emphasizes the critical effects of defining property boundaries; Smith

---

<sup>7</sup> Henry E. Smith, *Semicommon Property Rights and Scattering in the Open Fields*, 132 J. LEGAL STUD. 131, 132 (2000).

<sup>8</sup> Id. at 138–40.

gives the example of scattering individual farmers' holdings, so it would be harder for a shepherd to concentrate manure or grazing on any particular plot.<sup>9</sup> Third, it indicates the interplay between what Smith would later call the "exclusion" and "governance" strategies for monitoring resource usage as alternative institutions for preventing strategic behavior.<sup>10</sup> And fourth, notwithstanding the fact that monitoring and enforcement can be more difficult in a semicommons than in a purely private or a purely common system, there can be important synergies from enabling both forms of access simultaneously. The open fields worked best neither as pure pasturage nor as pure farms. *The Internet could not possibly work if it were entirely private or entirely common.*

## II. Moderation

### *Online Semicommons*

Smith's framework enables a more formal model of online community as semicommons.<sup>11</sup> There, the privately-held aspect of the resource is the underlying rival (and excludable) infrastructure, and the commonly-held aspect is the nonrival (and not naturally excludable) information goods users exchange. It's helpful to break users' activities into three roles: writing, reading what others have written, and engaging in moderation. The same person may of course engage in all three activities—as well as being an infrastructure owner.

- Authors may have some intrinsic motivations to write that are independent of the size of their audience, but they also value having a larger audience. For a given level of effort, an author's costs are independent of audience size.
- Moderators, like authors, care that particular items of interest to them be read by others. Moderators may disagree about which items are worth promoting. I will call a moderator's preferences among items her "ideology." Moderation, like writing, takes time and effort. Some of that effort is spent in reading enough to decide what moderation is necessary; some of it is spent actually flipping the necessary switches.

---

9 Id. at 161–64.

10 Henry E. Smith, *Exclusion and Governance in the Law of Nuisance*, 90 VA. L. REV. 965 (2004).

11 Robert Heverly has written of the "information semicommons," but he refers, instead, to the interplay between private and common uses of information. Copyrighted information is privately owned; the public domain is held in common. Robert Heverly, *The Information Semicommons*, 18 BERK. TECH. L.J. 1127, 1164–72 (2003). Heverly is thinking about all human use of information. My focus is narrower in that I look only at practices within a given online community, but broader in that I include the community's infrastructure as part of the relevant resource set.

- Readers gain value from reading particular items of interest and have highly idiosyncratic in their preferences across authors. Reading an item in full takes time and effort. By default, readers have no way of deciding which items to choose to read, but moderation can help them pick and choose.
- All three activities impose costs on the infrastructure, a rival resource that is costly to provision and could be used for other purposes. These costs are borne by the owners.<sup>12</sup>

An online community run as a semicommons can partake of several virtues. An ideal commons is cheap: the explicit and implicit costs of participation are zero. Users don't need to pay much (in money or effort) to use it, and the owners don't need to pay much to operate it.<sup>13</sup> On the other hand, the ideal commons is also productive: it generates valuable information goods. Authors reach large audiences; readers can find works they like; moderators approve of what readers can find; society benefits from positive spillovers. These virtues are economic, but others appeal to political values. An ideal commons is broadly open: the set of participants in the community should include as many as possible of those who would like to participate, as authors, readers, and moderators. And finally, an ideal commons is democratic: its users participate in the community's self-governance. These virtues are incomparable. By talking of virtues, plural, one can discuss a community without needing to decide, say, whether more democratic participation for members justifies greater restrictions on who can be a member. Interpersonal incomparability is particularly important; one community might place more costs on authors, another might place those costs on moderators. I hope to provide a useful vocabulary for others to debate whether some such choices are better than others, but such questions are beyond this paper's scope. For present purposes, the most important point will be that different forms of moderation make tradeoffs among these virtues in different ways.

---

<sup>12</sup> I am not including the limited time and attention of the community members in the privately-held portion of the communication semicommons. I model those effects in the participants' cost functions. I believe that these factors could be rolled into the description of the infrastructure without affecting anything essential in my argument; I have not done so because it seems unintuitive to talk about attention as a shared resource. Smith does not model the farmer's or shepherd's labor as part of the open-fields semicommons; I similarly think that participants' labor is not part of an online semicommons as such.

<sup>13</sup> If users and owners can profit, so much the better, but in assessing overall social benefit, it is easier to separate the operating costs from the benefits produced by the information goods, leaving wealth transfers (which balance out) to one side.

Following Smith, we are also led to analyze the ways in which users can impose costs on each other.

- In congestion, common users (principally authors) overuse the infrastructure, imposing provisioning costs on the private owner. If the use exceeds the infrastructure's capacity, costs also fall on other common users as messages are dropped or take longer to transmit.
- In cacophony, each author's contribution raises the search costs for readers and moderators by increasing the amount of material that they must sort through.<sup>14</sup> Spam is the example par excellence of both congestion and cacophony.
- In manipulation, ideological moderators impose costs on those who do not share their ideology. Common patterns of manipulation include edit wars (competing moderators exhaust the available rents from moderation), capture (private owners take over moderation), and deception (moderators lie to the community, e.g. through sock puppetry).
- In weaponization, authors distribute content of negative value: information bads. Common patterns of weaponization include harassment (intentionally targeting specific users), trolling and misuse (targeting the entire community), and causing externalities outside of the community (e.g. a web discussion board for encouraging the assassination of doctors performing abortions).
- In demoralization, other problems cause users and owners to stop participating. Demoralization is in a sense just the Tragic story from an *ex ante* point of view; those who do not expect the experiment to succeed will not take part in the first place.

These costs—the natural results of strategic behavior—also represent failures of specific virtues. Weaponization makes the commons less productive from society's point of view; demoralization can frustrate democracy. We shall see that different moderation strategies make different choices about how to prioritize virtues in the way they respond to different threats to the commons.

---

<sup>14</sup> For a given reader, or for readers in general, this effect may be offset by the benefits to readers of being able to read an author's contribution, or it may not. Much depends on where the contribution falls along readers' curve of diminishing marginal returns from further options.

## ***Verbs of Moderation***

The study of moderation strategies begins with the “verbs” of moderation; the basic actions that moderators and owners can take to affect the dynamics of an online semicommons. I start by identifying four key verbs: norm-setting, exclusion, pricing, and organization. Organization is a complex enough topic in its own right that I call out four subverbs: deletion, filtration, annotation, and synthesis.

### *Norm-Setting*

Shaping social norms is the first and most important task of moderation. If every user of a resource complies with a shared social norm of appropriate use, there is no further problem to be solved. Every other verb of moderation is in a sense secondary to norm-setting. Long-time community builders see their task as creating a shared sense of engagement; every other moderation lever is useful to the extent that it can build that sense, whether directly or indirectly.

Respectful norms cannot simply be set by fiat. By definition, they are an emergent property of social interactions. Individual participants, private owners, and system designers have limited power over group norms; most of the levers they can pull will only nudge norms in one direction or another, possibly unpredictably. Good norm-setting is a classic example of know-how; there are principles and known techniques, but knowing, say, whether to chastise an uncivil user publicly or privately is not a decision that can easily be reduced to a simple algorithm. Indeed, many such distinctions are significant only within a particular community. There are a few ways in which system designers and users can try to influence norms. Other users can model correct behavior; designers can state that certain behavior is or is not encouraged; and users can express disapproval of transgressors. More subtly, the system’s design may send a message; for example, the user interface could signal that certain contributions are particularly valued, or make it easier for users to send norm-reinforcing signals than to send norm-damaging ones.

### *Exclusion*

In an information semicommons, exclusion takes the form of prohibiting some individuals from participating. Rather than attempt to calibrate specific uses (“governance” is Smith’s term for this alternative), one simply excludes them from all uses. Exclusion can avert strategic behavior by removing negative-value content while keeping positive-value content, but in this balance, both Type I and Type II errors can be costly. Exclusion can work first through differential targeting, keeping out those users more likely to make abusive or excessive uses, and

second by limiting community size, since smaller communities may *eo ipso* have more desirable properties. The decision of who will be excluded can take place along a continuum from very precise (e.g. “Margo, Venkatesh, and Istvan, but not Oksana”), to very broad-brush (e.g. “all those with .edu email addresses, but not those with .com addresses”). At any level of precision, a particularly important decision is whether the default is inclusion or exclusion.

### *Pricing*

Some prices are explicit, such as a \$12.95/month subscription fee. Other prices are implicit, such as a ten-step signup process that takes twenty minutes to complete. Prices can even be negative; when it launched, Epinions paid users to write reviews. Pricing also involves a wealth transfer; one user’s price is another’s income stream. At the least granular level, one price purchases unlimited access to the entire community and to all roles. At the most granular level, each individual action is separately priced in distinct microtransactions.

The most common role of pricing is to make participants internalize some of the costs of their behavior. Pricing is more information-intensive than exclusion, however, in that one must also set the level of prices. Even imperfectly-calibrated prices can sometimes improve on an unpriced status quo; people might think more often before hitting reply-all if email cost a tenth of a cent per recipient. Indeed, many systems use clever interface design to make more cumbersome those actions more likely to be computationally expensive or annoying to others. The information cost involved in calculating prices is one of the characteristic savings of a commons system; one of the basic choices facing any online service is whether to be free or to charge. At extremely high prices, pricing collapses into exclusion as the price becomes prohibitive. Contrariwise, at extremely low prices, pricing becomes an invisible speed bump that eventually disappears entirely.

### *Organization*

Organization is any activity that affects the flow of contributions from authors to readers. It is the verb of moderation that best takes advantage of the informational capabilities of computers. Categorizing messages on a bulletin board by topic is organization; so is searching them by keyword; so is counting the number of total messages; so is deleting off-topic messages. These are all ways of remixing authors’ contributions to give readers a more satisfying experience. The Internet and Web 2.0 booms have given us thousands of implementations of novel patterns of organization. Their intricacies are only just beginning to be understood.

Four techniques of organization are notable. In deletion, the moderator deletes one or more contributions or otherwise prevents them from getting through. A spam filter is an obvious example of deletion; it delete emails that it believes to be spam. Deletion also has a nondestructive cousin: filtration, in which a reader is given a specialized view of a subset of contributions. A search engine is a filtration tool; it lets a reader see only contributions matching a particular query. In annotation, participants add additional information about contributions and contributors, to enable readers to select more intelligently from available options. eBay's feedback system annotates contributors; Metafilter's favorites system annotates contributions; In synthesis, participants combine pieces of others' contributions in a way that generates fresh content. Wikipedia is the ultimate example of synthesis; there, small and heterogeneous changes by individual users are synthesized into entire encyclopedia entries.

The most direct use of organization is to decrease cacophony by helping readers see only the content they desire. Filtration hides the low-value content; deletion removes low-value content outright; annotation enables readers to evaluate the content's relevance without actually reading it; synthesis turns many moderate-value contributions into one higher-value one. Organization has an indirect role in decreasing congestion; authors are less likely to produce infrastructure-hogging low-value content if they know that readers will not see it. It is a tool for manipulation in the hands of self-interested moderators, but it can also provide tools for readers to detect and correct for moderators' biases. And, depending on how it is used, it can either greatly amplify or greatly inhibit weaponization.

One can use multiple techniques in parallel. An email list moderator who deletes some posts and flags others as "important" is both filtering and annotating. One can also use the output of one technique as the input to another. Amazon's average customer review ratings provide a synthesis of the ratings given by each reviewer; those ratings are in turn self-assigned annotations of the longer textual reviews. Similarly, organization interacts extensively with the other verbs, especially norm-setting. eBay has strong social norms against giving negative feedback except in serious cases of misconduct. In many communities, those who are flagged by other participants for poor contributions may be banned (exclusion based on annotation based on social norms).

### ***Characteristics of Moderation***

Picking a verb of moderation does not end the process. Each verb can be used in quite different ways. There are four important dichotomies in moderation: (1) humans vs. computers,

(2) secret vs. transparent, (3) *ex ante* vs. *ex post*, and (4) centralized vs. decentralized. These age-old categories capture some of the deepest tensions inherent in any system of governance, so it should be no surprise that they also capture tensions inherent in online governance. In an online semicommons, these four decisions are independent; any verb of moderation can be applied using any of the sixteen possible combinations. For example, spam filters are a secret decentralized computer *ex post* form of organization (specifically, deletion); a chat room facilitator is a centralized human transparent norm-setter who acts both *ex ante* and *ex post* and may have access to tools for exclusion and deletion. Describing moderation using these four dichotomies picks up many of the important characteristics of particular moderation strategies.

#### *Automatization*

Moderation decisions could be made by software or by people.<sup>15</sup> A policy against abusive language, for example, could be implemented either through a software filter or by a moderator decides what does and does not cross the line. Software is capable of engaging in any of the four verbs of moderation, albeit with different aptitude: it's best at pricing and exclusion, getting better at organization, and decidedly weak at norm-setting. Software moderation usually has higher fixed costs and lower marginal costs than human moderation. Thus, at very high volumes, software moderation is much cheaper than human moderation. Software is also better at adhering to a rule than humans are but less capable of exercising discretion well. Thus, software is comparatively more effective at making decisions that can be reduced to “hard” facts and figures, such as how many messages a user has sent or how widely a given message has been distributed.<sup>16</sup> Software is also generally more vulnerable to hacking and reverse engineering than humans are.

---

15 In one sense, the “modality of regulation” in all moderation is software; an online semicommons is a creature of software, and all decisions about its use are mediated by software. In another sense, all policy decisions about the use of a semicommons are made by the people who design and use the code. I am concerned with the intermediate question of which actor is responsible for day-to-day, garden-variety moderation decisions.

16 In Smith's framework of exclusion and governance in the semicommons, computers may have lower costs than humans for monitoring very specific classes of norm compliance, but higher costs where the norms are more context-dependent. The availability of computer enforcement creates new options both for exclusion and for governance.

### *Transparency*

We have already seen that moderation can affect norms by making some contributions more or less visible. This point also applies to moderation itself. Publicly acting against harmful messages can be just as bad as letting them get through; trolls will act to provoke a visible counterreaction, and participants will still be reminded that the trolls are active. The effect may be to weaken compliance norms, just as the IRS's annual tax-time saber-rattling about anti-evasion efforts makes some otherwise honest taxpayers aware that others are playing them for suckers. But secrecy can be just as bad. It can suggest that moderators have something to hide, or do not value participation; failed attempts at censoring harmful content can produce anti-moderation backlashes. More cheerfully, even ineffective moderation can sometimes be successful at setting norms, provided it gives participants the right sense that they are engaged in an enterprise where most participants cooperate and rare cases of defection are appropriately handled.

The choice between secrecy and transparency interacts with the choice between software and people.<sup>17</sup> Software is by default secretive; explanations must be coded in (at some additional cost) in advance. On the other hand, once explanations are part of the software, they are inexpensive to produce. The marginal case-by-cases costs of human-supplied explanations can be quite high; moderation decisions can be quite granular, while explanation may involve significant writing. The difference between a mouse click and a sentence's worth of typing can add up rapidly. Secrecy can sometimes (though hardly always) be a way of preventing a software rule from circumvented by making reverse engineering harder.

### *Moderation Timing*

Moderators can act *ex ante*, using their power over the infrastructure to allow certain actions and prohibit others. Or they can act *ex post*, using their powers to punish evildoers and set right that which has gone wrong. The decision to act *ex ante* or *ex post* is independent of whether humans or computers make the decisions; for example, a rule that all posts must be associate with one of a limited set of topics on a discussion board could be enforced by requiring human

---

<sup>17</sup> The debate between openness and secrecy in computer security is rich and the principles are not always easy to master. Openness can encourage both benevolent and malicious parties to examine software closely; secrecy can both slow down intrusion and slow down the detection of that intrusion. One commonly-cited principle holds (roughly) that most details about how a computer system operates should be public, except for a small central core that is kept tightly guarded.

authors to engage in *ex ante* annotation and pick a topic, or by having a computer categorize all uncategorized messages *ex post* once an hour. The *ex ante* / *ex post* distinction plays out differently for different verbs.

- Exclusion *ex ante* involves screening to predict which users will behave well and which poorly. (“We only invite new users when two current users vouch for them.”) Exclusion *ex post*, on the other hand, is a punishment for misbehavior. (“You have been banned for making insensitive remarks towards other users.”) Exclusion as a punishment can be mild (a temporary suspension) or severe (a permanent expulsion). The threat of retrospective punishment can have an *in terrorem* effect in convincing participants to moderate their own behavior.
- *Ex ante* pricing implements the usual jurisprudential understanding of a market: pay to play. *Ex post* pricing, on the other hand, could be one of two things. It could be “honor system” pricing, as with a shareware program, or it could be a punishment for misbehavior, what Robert Cooter would call a “sanction” instead of a “price.”<sup>18</sup>
- *Ex post* social norms involve the community expressing its approval or disapproval once the participant has acted; *ex ante* social norms are those that participants have internalized. Most norms are a little bit of each.
- The choice between *ex ante* and *ex post* organization is tied to the choice of actor. Authors act *ex ante*; readers act *ex post*; moderators can do either. The choice is therefore bound up with the distribution of implicit costs. Generally, *ex post* organization is a bread-and-butter daily task of moderators; the job of cleaning things up and repairing minor mistakes before they contribute to worse ones creates a great many tasks that are individually small but collectively important to a healthy sense of order. Regardless of who performs it, *ex ante* organization has a time cost; content does not appear until it has been reviewed. This delay can inhibit positive feedback loops.

#### *Centralization of Moderation*

Moderation decisions could be made either centrally by a single moderator whose decision affects the entire semicommons, or by multiple moderators whose individual decisions

---

<sup>18</sup> Robert D. Cooter, *Sanctions and Prices*, 84 COLUM. L. REV. 1523 (1984).

affect only part of the semicommons.<sup>19</sup> There are, of course, many gradations and many hybrid forms, and most systems are centrally moderated in some ways and not in others. Centralized moderation can act on the basis of a huge amount of coordinated information; decentralized moderation can act on the basis of hard-to-aggregate local knowledge. Centralized moderation provides global consistency and Comedic scale effects; decentralized moderation promotes flexibility and local resistance to Tragic invasion. Centralized moderation offers the ability to stop unwanted content and participants by creating a single checkpoint through which all must pass; but since one person's spam filter is another person's censorware, decentralized moderation can permit those with ideological differences to agree to disagree.

### ***Community Characteristics***

Just as one size does not fit all forms of moderation, one size does not fit all communities.. Four properties of a community are particularly salient in affecting the kinds of strategic behavior threatening it and the effectiveness of various patterns of moderation: (1) the capacity of the infrastructure, (2) the size of the user community, (3) the distribution of ownership, and (4) the identifiability of participants. As above, these characteristics are independent from each other.

#### *Infrastructure Capacity*

Use of the common aspect of the semicommons places costs on the private portion by consuming some of its capacities: storage, bandwidth, and CPU time. Bandwidth and CPU time automatically regenerate; storage can only be reclaimed by deleting data. In general, any given individual use consumes only a tiny amount of these three capacities; the impact is only significant when a great many uses are aggregated. All three are commodities, so there is a

roughly linear relationship between cost and capacity.<sup>20</sup> Capacity is, however, sticky, because the units of provision are lumpy, because adding or removing infrastructure is not instantaneous, and because there are certain natural inflection points as one shifts between technologies (e.g. from renting server space to collocating entire servers) in the same way that different gears in a car are appropriate at different speeds. This stickiness introduces a significant short-run discontinuity. Up through some critical value, additional usage has no effect on infrastructure performance; beyond that value, performance degrades rapidly. An online semicommons in which capacity is a significant bottleneck looks very different from one in which it is not. With little capacity, the Tragic aspects of congestion are prominent; as capacity increases, congestion problems recede and social factors come to the fore. A particularly interesting case of abundance arises from the lumpiness of computing resources. The minimum practical unit of provision is often sufficient to enable a great deal of usage. If ownership is widely distributed, this lumpiness may suffice to provision a great deal of capacity.

### *Community Size*

A closely related issue is the number of users. From the Tragic-Comedic point of view, the size of the community is *the* pivotal issue; whether common use generalizes to large groups is the heart of their disagreement. Increasing the community size has two offsetting effects for readers and authors. On the one hand, it catalyzes Comedic network effects; on the other, it generates cacophonous congestion effects for readers. Growth is also often unkind to social norms. As community grows, it becomes easier for individuals and groups to resist a norm. This breakdown makes it harder to use social norms to moderate in large communities. One common response is to embrace decentralized moderation, thereby fragmenting the community into smaller groups that can maintain their own norms. Organization, however, can often benefit from size, by way of the law of large numbers. Any individual moderator's assessment of an action's value may or may not accurately reflect the community's sense of value, but the average of a thousand moderators' assessments is likely to express it fairly well. Pricing at large volume can benefit from the salami-slicing effect; a great many small payments can add up to a surprisingly

---

<sup>20</sup> There are economies of scale in provisioning all three, because of declining setup and labor costs involved in running large server farms of identical components. Some of these economies are partly offset by the engineering problems involved. You cannot make a petabyte database simply by connecting a thousand terabyte databases to each other.

large number. Finally, community size shapes the ways in which moderation can best be carried out. All the verbs have costs that increase with volume (whether human or automatic); the second-order problem of provisioning moderation requires greater and greater investments as a community grows.

Community size and infrastructure capacity together give rise to a number of scale transitions familiar to those who work with online communities:

- The transition from beneath the size at which a single owner can self-provision all needed capacity from the lumpy unit she needs to participate as a user.
- The transition from too few participants to sustain a conversation to enough.
- The transition from a community size that can sustain cooperative social norms without any coercive help to a group large enough for some participants not to care what others think of them.
- The transition from a community small enough to be moderated by one person without help to one that can't.
- The transition from a community small enough to be moderated by a centralized group of humans without software support to one that requires automated tools or decentralization.

#### *Centralization of Ownership*

Instead of a single owner, an online semicommons could have many, each of whom owns a piece of the infrastructure. This is the pattern described by Smith in his discussion of scattering. Distributed ownership in an online semicommons can have a similar function: helping align incentives by making the heaviest users of a particular piece of infrastructure its owners. But centralized ownership also has benefits. It can, for example, be a counter to weaponizing strategic behavior in which users deliberately try to drive up a private owner's costs by targeting her infrastructure for heavy use. Email bombing and distributed denial-of-service attacks take advantage of distributed ownership to censor a participant by overwhelming her piece of the infrastructure. The choice to concentrate or disperse ownership also affects the political economy of the choice among moderation patterns. Owners can use their power over the infrastructure layer to make policy at the content layer. If moderation is centralized, then as in corporate governance, concentrated ownership enables closer monitoring to make sure that moderation is effective. On the other hand, centralized ownership raises the danger that the owner will capture

the moderation to further her own interests. Users who are also owners may be better able to protect their interests by demanding a say in moderation policy.

### *Identifiability*

At one extreme, users of an online semicommons could bring with them a complete biography of online and offline lives. At the other, they could have no stable identity at all; many online polls are totally and completely anonymous, to the point that a participant could vote more than once because the system does not even divide the world into those who have voted and those who have not. There are many gradations in between. The most important role of identity is creating stability through time, so that others can link past behavior to a present identity. All four verbs of moderation can tap into this continuity. Exclusion absolutely depends on it; without identity, the distinction between “outsiders” and “insiders” would not exist. In theory, pricing need not build on any persistent identity; in practice, any explicit pricing system for an online resource will depend on some non-trivial infrastructure of identity, such as a credit card processor. Identity is often vitally important to norm-setting. The online comic strip *Penny Arcade* explains why so many gamers behave crudely in online games with the equation “Normal Person + Anonymity + Audience = Total Fuckwad.” Creating stronger persistent identity can make positive social norms more effective; at its best, it creates the conditions for the effective monitoring and graduated sanctions beloved by common-pool resource scholars. On the other hand, strong identity also sometimes creates a badge for misbehavior. Organization can both piggyback on and produce identity. Filtering and deletion both often treat the identity of an author as a significant data point. In reputation systems, participants provide annotations on each others’ actions, and those annotations become part of one’s community identity.

Paradoxically, both identity and its opposite—anonymity—can be expensive to establish. Externally-produced identity requires participants to prove facts about themselves, which can cost both time and money; internally-produced identity requires participants to take the time to learn about, comment on, and rate each other. Anonymity might seem cheaper, but genuinely hiding identity so that the curious will be unable to learn it requires some significant effort—deleting log files, stripping out snooping software, and taking action against participants who “out” one another’s offline identities. And finally, identity can be the enemy of privacy, for good and for bad. Divulging information about oneself is itself a cost.

### ***Democracy and Meta-Moderation***

The last question of moderation may be as large as all the others put together. Democratic participation by users in setting moderation policies is not only a virtue in itself, but also a significant factor in determining the moderation patterns in use and the effectiveness of those patterns. The first-order description of governance in an online semicommons is trivial: those who control the infrastructure are the leaders, and their software-enforced decisions are binding.<sup>21</sup> But this is hardly a complete description; a wealth of formal and informal mechanisms are available by which the infrastructure owners may (and often do) consult participants or delegate extensive moderation powers to them. There are online communities that are governed as dictatorships, as consultative benevolent dictatorships, by consensus among a group of self-selected administrators, by straw poll, by formal legally-mandated democracy, and by Hobbesian anarchic scrum.

Democratic institutions can be a powerful weapon against capture of a semicommons by one faction or viewpoint, but they also can be the instrument of that capture. There are online communities in which the private owner runs roughshod over users' wishes; there are also communities where entrenched long-time users actively seek to drive out newer members who might threaten their influence. A particularly difficult problem for online communities with open borders is that defining the scope of the community entitled to representation and to be heard becomes difficult. Relying on exit and assortative matching is problematic for the Comedic goal of creating one unified supercommons.

Democratic processes also affect the available moderation patterns. Anyone can participate in setting rules for human moderators to implement, but participation in designing software moderation is possible only if the community includes someone with the ability and willingness to make appropriate software changes. This fact leads to substantial technological lock-in in many communities; the software system they are using can be tweaked but not substantially changed. Democratic processes connected to particular decisions about moderation may be incompatible with some kinds of secrecy in making those decisions; this fact can be seen

---

<sup>21</sup> Norm-setting is the sole exception to this point. Any participant capable of sending any sort of message to others can participate in norm-formation. Excluding participants from this process may often be impossible; they can communicate through channels outside of the semicommons itself (e.g. many virtual world players discuss their worlds from outside, on extensive web discussion boards).

either as a constraint on how secretive a fair decision-making process can be, or as a constraint on how much public discussion is possible around a secret but effective decision-making process. Decentralized ownership creates a second, parallel constituency whose interests the structures of participation must accommodate; decentralized moderation decreases the scale on which individual structures may need to operate. Without some level of stable participant identity, similarly, certain kinds of democratic structures (e.g. binding plebiscites by majority vote) are impossible.

### **III. Case Studies**

The previous Part argued that online communities have available a wide range of moderation patterns to protect commons virtues from strategic behavior—but that each pattern comes with its own distinctive costs for those same virtues. This Part will analyze those patterns in action in several case studies.

#### ***Metafilter and Slashdot: Multiple Paths to Success***

Our tour begins with two web sites dedicated to sharing interesting links. Metafilter (“community weblog”) and Slashdot (“News for nerds. Stuff that matters.”) face roughly analogous problems and have roughly analogous structure, but solve their moderation problems in remarkably different ways.. Both are structured around a set of user-submitted “front page posts” (for Metafilter) or “stories” (for Slashdot): links to something interesting on the web, along with a few sentences of description. Each story has an accompanying discussion page, on which users post their own commentary and conversation sparked by the initial link. Slashdot has about 25 stories a day, typically with a few hundred comments each, and about 150,000 registered users. Metafilter also has about 25 stories daily, typically with 20-100 comments each, and about 50,000 registered users. More people read both sites than post to them; Metafilter sees about 3 million unique IP addresses s a month; Slashdot, about 7 million. Both communities are mid-sized by Internet standards, and orders of magnitude larger than the size at which one normally expects cooperative norms to start breaking down.

Slashdot filters all story submissions through a small (half a dozen) set of moderators, who choose which stories will appear. The number of submitted stories is small—a couple of hundred a day—which makes this human centralization tenable. Comments, on the other hand, are wide open. Both registered users and “Anonymous Cowards” can post comments immediately. The need to associate a comment with a story provides a first level of structure, and comments can be

threaded to create conversations. Registered users can “moderate” other users’ comments, giving them a +1 or -1 ranking for various reasons, such as “interesting,” “insightful,” “offtopic,” or “troll.” These scores are summed on a range from -1 to 5; readers can then choose to see only those comments above a certain threshold. Slashdot quickly discovered that moderation could also be abused, and has instituted increasingly baroque systems to watch the watchmen. In “meta-moderation,” registered users judge whether randomly-selected moderation decisions were justified or not. These meta-moderation decisions then become feedback into a “karma” system under which only those moderators who use their powers for good rather than evil (as judged by a complex but open-source computer algorithm) are allowed to moderate in the future.

The moderation/metamoderation/karma system is the core of Slashdot’s self-regulation, but there are other systems at play. IP addresses that try to engage in denial-of-service attacks or post malicious comments will be banned. There is an experimental user-supplied tagging system. Comments with malformed HTML (which could create security holes or rendering problems) are blocked by software and will be removed on sight by the site admins if discovered. The site shows advertising, but readers can purchase subscription that give them a thousand ad-free pageviews for \$5. Slashdot’s norms are particularly interesting. The site has an explicit technolibertarian philosophy; its administrators do not delete posts for reasons of content and they have fostered a climate of free-wheeling high-intensity exchange, in which tempers frequently flare and insults fly. The site has had an extensive history of deliberate misbehavior; abusive users would attempt to score a “first post” by jumping on a newly-posted story, horrify other readers by displaying disgusting images such as the infamous goatse, and troll other users by posting obvious (in hindsight) flamebait. In fact, flagrant misbehavior of this sort is still regularly attempted but is now routinely moderated down and out of sight (most users set their comment filters to 0 or above, so that the -1 posts simply disappear from view). Slashdot discussions range from the sophisticated to the sophomoric, but even a casual browse reveals that the moderation system is effective in screening out most of the true dross.

I would not say that Metafilter could not be more different from Slashdot—it could—but the differences are nonetheless striking. Where Slashdot membership is open to anyone immediately; Metafilter requires a \$5 fee to sign up. (During times of crisis, such as when a group of outsiders is flooding in, new signups are sometimes disabled.) Where Slashdot filters posts through moderators, Metafilter lets anyone with an account create an FPP. Comments are

organized by FPP, but otherwise are unthreaded. Users can mark their favorite FPPs and comments, and also flag FPPs and comments for removal. Favorites are displayed publicly, while flagged posts are shown only to the site administrators. The three administrators delete posts and comments for various reasons, usually with a short standard explanation, and paying special attention to repeatedly-flagged ones. There is a norm-setting asymmetry here; positive contributions are praised and made more visible, while negative ones are hidden. Matt Haughey, the site's founder and lead administrator, also maintains a sideblog on the front page of "new and noteworthy posts" that also highlights particularly excellent contributions. Users who behave disruptively will draw disapproving comments from other users, then gentle emailed reminders from Haughey, then more serious warnings, and finally an outright ban. The site is advertising-supported; Haughey has cycled through a number of different advertising systems, all of them unobtrusive.

Haughey and his right-hand co-admin, Jessamyn West, are absolutely certain that the secret to Metafilter's success is generating positive, cooperative community spirit—in my terminology, norm-setting. They think of all of their moderation actions in terms of how they will affect group norms. Thus, they are heavy and visible site users themselves (Haughey credits the site's initial takeoff to the weeks during which he was the primary author of FPPs and active in all discussion threads, setting an example for others) and heavy email correspondents, offering pats on the back and gentle reproaches. Offshoot sites provide additional opportunities for users to blow off steam, provide feedback on the site, and learn about each other: particularly MetaTalk, where they discuss the site itself. They are willing to explain any action to death, try very hard to keep their own emotions out of their actions, and move quickly to push back against actions that threaten shared norms of cooperation. Partly as a result, Metafilter has a solid (if constantly changing) core of about 300 heavy users who reinforce norms of basic respectfulness and self-restraint, and a longer tail of thousands more who participate on these largely positive terms. There have been a string of crises over the years—real disputes among camps of members, invasions by crowds unfamiliar with Metafilter but angered by something posted there, near misses with upset lawyers, and occasional harassment—but the site has successfully weathered multiple scale transitions with a basic sense of intellectual ferment roughly intact.

Metafilter and Slashdot's experiences have some important similarities. Both are run by administrators who are deeply involved in day-to-day conversations on the site, who rely on their

own participation and extensive statistical instrumentation to keep in view what is happening. Neither is at all a democracy in a formal sense, in that the administrators on both set policy unilaterally. In West's words, "At the end of the day, someone always has root; ignore this at your peril." Both now rely on core groups of participants; Slashdot has a smallish core of regular story contributors, while Metafilter has its version of the 300. Both produce a diet of links and a much larger flow of conversations, both of which are thrown open to the wider world. But the richness of moderation is such that they achieve their success using remarkably different strategies. Slashdot's moderation/metamoderation system is a complex system of aggregating user feedback on other users in an algorithmic way; Metafilter's "system" is really a rich set of norms, in-jokes, customs, habits, and styles. Slashdot is stable because many users want it to work and because its algorithms have been well-tailored to provide a strong lock-in that prevents a few users from disrupting the organizational machinery. Metafilter is stable because it has a critical mass of dedicated users who apply lesser social sanctions, who are backed up by conscientious and active administrators who can apply stronger sanctions as needed.

### ***USENET and Email: Not All Moderation Succeeds***

USENET newsgroups (started in 1979) and email (1982, based on protocols dating back to 1971) are both old technologies, by Internet standards. They have roughly similar overall structure; both are systems that allow users on computers across the Internet to communicate through a peer-to-peer process of message exchange with local storage. Email prospered on the massive scale-up of the 1990s Internet boom; USENET was overwhelmed by that same boom. Their divergent paths illustrate how the choice among moderation strategies can doom or save a communication semicommons.

USENET is a distributed set of message boards, with a peer-to-peer protocol by which different sites exchange new messages. Each server talks only to a few others, but most USENET servers are linked together so that any given message will eventually be propagated to all servers in the network. In 1987, participating sites agreed to a "Great Renaming," establishing a canonical hierarchy of topical newsgroups (such as rec.pets.cats, sci.math, and alt.fan.karl-malden.nose); a given message is posted to one or more newsgroups. Some groups require the approval of a moderator to post a message; others do not. Post can be deleted by the author; these "cancel" messages are easily forged, so site administrators often set their own policies on whether to honor cancel requests. Many users use "killfiles": lists of other users whose posts will

be hidden from the killfiler. Creation and deletion of newsgroups normally goes through a deliberative voting process coordinated by a centralized board that publishes a canonical list of approved newsgroups. In the alt.- portion of the hierarchy, however, site administrators simply decide whether or not to add a new group, and other administrators decide whether to follow their lead.

This abbreviated description of the system reveals many familiar organizational techniques, including annotation (choice of newsgroup), filtration (killfiles), and deletion (cancels of others' posts). The hierarchy system has centralizing democratic procedures almost hardwired into it, but also respects decentralized values with the less-constrained alt.- hierarchy. The peer-to-peer setup gives it good scaling properties, since sites bear much of the costs of hosting and communicating created by their usage. The division into different newsgroups also generated strong group community norms within different newsgroups; people drawn together by shared interests have often developed strong community feeling.

For all of these virtues, USENET's moderation patterns did not deal well with the flood of new users who came online in the 1990s. In 1993, AOL began offering its subscribers access to USENET newsgroups, an event known as the "Eternal September"—a neverending stream of new users as unfamiliar with USENET's norms as the annual crop of college first-years had been. In 1994, the commercial spammers arrived, starting fittingly enough with a pair of lawyers who cross-posted an advertisement for their immigration services to thousands of newsgroups. Massive cross-posting, combined with a lack of respect for norms of self-restraint, became endemic. Griefers made both targeted attacks on particular newsgroups and general attacks on hundreds at once; when an unsophisticated user would try to respond to insist that they leave, she would often simply end up cross-posting herself, adding to the chaos. Vigilante USENET users began automating cancel requests to counter automated cross-posts, but the damage to many previously healthy newsgroups was fatal. As web-based discussion boards and third-party hosted mailing lists became increasingly reasonable alternatives, many of the conversations that had taken place on USENET shifted elsewhere or disappeared. USENET itself is not dead—one can still go to Google Groups or Giganews and participate in ongoing conversations in groups with strong norms that allowed them to weather the storm—but it has nowhere near the relative importance that it once did to the life of the Internet. USENET fell victim to the moderation

equivalent of the Peter Principle; it grew until its moderation mechanisms could no longer cope with the scaled-up disruptive behavior, and there it stopped.

The email system did not stop growing in the mid-1990s, even though it, too, was deluged with spam and (to a lesser extent) with denial-of-service flood attacks. USENET's Achilles heel was that it had centralized organizational patterns but no effective system by which users could pool their efforts to defend that organization from large-scale attack. Each newsgroup spanned all servers; each server hosted all newsgroups. Unlike USENET's NNTP, SMTP is not a replication protocol, but a transfer protocol. Emails go to their recipients, full stop. There are no email equivalents to newsgroups—coordinated entities that all users see in a substantially identical form. Each email server is a dedicated piece of infrastructure designed to enable incoming and outgoing email for its own users. I can install a spam filter without disrupting any email except that to and from users on my piece of the network. This difference has provided for greater flexibility in experimentation with local anti-spam policies. Spam remains a serious, expensive problem, but email remains usable and valuable. The network of regular email users continues to grow.

The tradeoffs between freedom and control in the email system are fascinating. Email scores highly on many virtues. Access to email is exceedingly broad, and that access is close to free for users. Once one has an Internet connection, one can usually get free email from one's ISP or ad-supported email from a webmail provider. Individual users have substantial control over their own spam filters, a little say in their ISP's spam filters, and almost no say in any global properties of the email system (which has little in the way of formal governance). Individual-to-individual email is mostly uncensored and uncensorable by other users. But all of these virtues come at the price of some specific design tradeoffs. Email is deeply non-public; a great email message can only be shared with many people through successive forwarding, thereby foregoing some productivity. The decentralized system that makes spam locally filterable also makes it hard to stamp out spam on the sending side, creating large back-end costs and moderate costs for users who must tend to their spam filters. The inconsistent and decentralized system of anti-spam enforcement also creates unaccountable, sometimes undetectable drop-outs in email connection—messages sometimes are thrown away without warning and for reasons that may be hard to fathom. There are regular calls to make email more accountable—making senders pay to send messages, creating stronger authentication, or establishing central anti-spam systems.

Centralization of these functions, however would raise difficult democratic issues of participation and accountability, would threaten users' abilities to experiment different local spam policies, and would threaten the live-and-let-live approach that allows users with violently different beliefs all to use email.

### ***Wikipedia: Dynamic Evolution***

Wikipedia is a favorite example of Comedic theorists. Its rapid growth and surprising reliability make it a powerful proof-by-example of the power of large-scale Comedic sharing. I do not plan to rehash what has already been written about Wikipedia's sudden success. Instead, I would like to focus on an underappreciated aspect of Wikipedia's moderation policies: their dynamic evolution. Wikipedia has thrived in part because its moderation policies have adapted to changing challenges.

The founding moment of Wikipedia is usually considered to be Jimmy Wales and Larry Sanger's January 2001 decision to augment the peer-reviewed Nupedia with a wiki open to anyone to edit. This move shifted from a system with substantial exclusion barriers for author to one without, and also to one with a much less (implicitly) expensive contribution process. This new process was much more friendly to rapid large-scale editing, and catalyzed rapid growth both in the number of articles and in the number of contributors. (Wikipedia relies on synthesis in a deep way, allowing editors to make incremental improvements on previous contributions.) The success of these early efforts, in turn, had a useful norm-setting effect in drawing more contributors in. It was, in short, a virtuous cycle. And at first, page vandals could be kept in check simply by the regular efforts of conscientious editors. The subsequent history of Wikipedia is also a history of innovative responses to threats. Here are three interesting examples:

*Protection and Blocking:* Wikipedia's growth has made it attractive both to vandals and to ideologues. Users who engaged in repeated or large-scale vandalism, who take part in vicious edit wars and will not back off to discussion when asked to, who target other users, or who flout other important community polices are now subject to IP-addressed based blocking, a form of ex post exclusion. This policy is exercised in tandem with protection, under which controversial or important pages at a high risk of vandalism or ideological edit wars are "protected" with varying degrees of intensity from edits. Full protection prevents everyone other than administrators from editing a page; semi-protection prevents new users from editing pages. These are forms of ex

ante deletion. Protection is regularly used not merely to prevent norm-defying users from making changes, but also to reassert norms through cooling-off periods.

*Dispute Resolution:* Wikipedia's reliance on centralized synthesis gives it a classic unavoidable problem common to such schemes: resolving disputes over which view will prevail. To an extent that should not be underestimated, Wikipedia often simply ducks these issues. Many Wikipedia entries adopt a measured "on the one hand . . . on the other hand . . ." tone towards many questions, legitimating some quite questionable views in an excess of caution. This response, as frustrating as it can sometimes be in a would-be authoritative encyclopedia, is an understandable technique of community maintenance that avoids alienating participants by rejecting their views. Wikipedia also uses a confusingly wide array of dispute-resolution techniques, including taking votes on article-deletion requests and a system of tiers of review for policy decisions. These mechanisms are largely self-generated, an outpouring of cooperative creativity that would make Ostrom proud.

*Participation-Based Identity:* Having abandoned Nupedia's system of pre-credentialling, Wikipedia is faced with difficult identity-related challenges. Matters such as page deletion are conducted by vote among those interested, making sock puppetry attractive to those who would win a dispute. More generally, in a large community where any user has the technical power to alter the outputs, some system by which users can sort out who else to trust in case of dispute is essential.<sup>22</sup> This system of trust is social; long-time Wikipedians who have made many edits, participated in many discussions, and been praised by many other Wikipedians have greater socially-granted authority. This measure of community respect is used in determining which Wikipedians will become administrators with greater code-based powers. It is also used as a sign of authority in disputes all across the site. Long-time Wikipedians know about each other and will often stick together to reassert Wikipedia norms and internal operating rules. This social coherence has provided legitimacy for many organizational projects and procedural standards that give Wikipedia and Wikipedia editing a somewhat predictable structure.

---

<sup>22</sup> Another important prerequisite for stability is part of any wiki software: keeping page histories so that any mistaken edit can be undone, or "reverted." No user actually has the power to delete previous contributions. (Deleting entire pages is a difficult border case because the deletion removes the page history. Unsurprisingly, Wikipedia has special policies on point.)

These mechanisms all have costs. Protection inhibits the editing of important pages; disputed topics are sometimes simply frozen as is for the sake of stability. The dispute resolution systems are often ad hoc and confusing. And the informal credentialing system favors long-time contributors. All three mechanisms have played roles in helping Wikipedia grow and maintain coherence against directed attacks and emergent chaos. Especially taken together, however, they create an increasing risk of capture—newcomers are presented with a set of complicated and inconsistent community rules, which they are then penalized for not knowing. Prizing community contribution over intellectual contribution is a natural pattern to adopt. But for an encyclopedia, it has substantial risks—including an open disdain for expertise, a sometimes comical lack of interest in contributor honesty, and hostility to new users. The communal social experience of Wikipedians both sustains and threatens Wikipedia as a knowledge-generating institution; it is both an instrument of freedom and of control; it both deters and constitutes strategic behavior. Wikipedia's ability to develop these institutions is a testament to the power of its deliberatively democratic processes and the responsiveness of those with control over the code to the needs of the community. The most positive indicator for Wikipedia's ability to continue its success is not that its moderation patterns currently work, but that it has developed moderation patterns dynamically in response to past challenges.

### **Conclusion**

This paper began by noting the debate over whether notable successes of commons production thus far, such as open source software and Slashdot, are scaleable, generalizable, and sustainable. Comedic theorists emphasize the productive power of freedom; Tragic theorists emphasize the need for controls to prevent waste. The theory of moderation I have presented emphasizes that freedom and control are not absolutes. No community is every wholly free or wholly controlled. Its patterns of moderation place it somewhere in between. To say that a community must adopt new patterns of moderation does not require one also to say that the community thereby becomes unfree, or is no longer meaningfully a commons. Thus, while Slashdot might experience an influx of users less sympathetic to its shared norms, and might thereby need to respond with new layers of moderation, it does not follow that Slashdot will score poorly on the checklist of commons virtues. Everything depends on the dynamic equilibrium of moderation and threats.

Moreover, the diversity of moderation patterns and their applicability to a wide variety of situations provide a reason for mild optimism about long-term prospects. Moderation has shown a tenacious ability to adapt to new challenges, as the example of Wikipedia shows. Moderation may be controversial, but it works. I suspect that a community's ability to adopt a successful set of moderation patterns is far more dependent on political factors than on the availability of a set of moderation patterns that would do the job well. This claim is, in a sense, just a modified version of Ostrom's point about when cooperative commons governance succeeds and fails. Moving to an online environment offers a richer set of possible institutions, because the community may have control over software, offering an ability to change the rules of the game usually absent offline. Future work on the sustainability of online semicommons and commons production will need to draw upon the rich interdisciplinary literature on the political development of institutions. The present theory of moderation combined with a good theory of online community politics should offer useful predictions about when a semicommons will be able to adapt its moderation to deal with new problems, and when it will not.

Moderation is an essential missing term in current discussions of the commons. Good moderation can make a community thrive; bad moderation can doom it. It balances freedom and control; it trades off virtues against one another. At its best, it enables Comedy without also opening the door to Tragedy. Moderation is a tool whose subtlety, power, and complexity we are only just beginning to appreciate.