Unraveling Internet Identities:
Accountability & Anonymity at the Application Layer

by

Josephine Charlotte Paulina Wolff

A.B., Mathematics, Princeton University (2010)

Submitted to the Engineering Systems Division
in partial fulfillment of the requirements for the degree of

Master of Science in Technology and Policy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2012

Author ....
Engineering Systems Division
May 11, 2012

Certified by ....
David D. Clark
Senior Research Scientist
Computer Science and Artificial Intelligence Laboratory
Thesis Supervisor

Accepted by ....
Joel P. Clark
Professor of Materials Systems and Engineering Systems
Acting Director, Technology and Policy Program

Unraveling Internet Identities: Accountability & Anonymity at the Application Layer

by

Josephine Charlotte Paulina Wolff

Submitted to the Engineering Systems Division
on May 11, 2012, in partial fulfillment of the
requirements for the degree of
Master of Science in Technology and Policy

Abstract

Both anonymity and accountability play crucial roles in sustaining the Internet's functionality, however there is a common misconception that increasing the anonymity of Internet identities necessitates diminishing their accountability, and vice-versa. This thesis argues that by implementing accountability mechanisms and anonymity protections at the application layer of the Internet, rather than the network layer, it is possible to develop a variety of different types of accountable-anonymous virtual identities tailored to meet the needs of the great diversity of online applications. Examples are drawn from case studies of several identity mechanisms used by existing applications, including e-mail, the virtual community Second Life, the Facebook social network, and the review site Yelp. These case studies focus on potential "points of control" for each application, as well as the ways different proposed identity schemes can leverage these control points to help mitigate the problems observed in existing identity frameworks, especially the issue of "discardable identities," or online identities that can be easily and cheaply disposed of and replaced. General design patterns for implementing accountability are discussed, with particular emphasis on the design of application-specific identity investment-privilege trade-offs, conditional anonymity schemes, and aggregated, identity management systems, as well as the role of scoped identities and linked identities in promoting online accountability.

Thesis Supervisor: David D. Clark
Title: Senior Research Scientist
Computer Science and Artificial Intelligence Laboratory

Acknowledgments

Returning to Cambridge after a four-year hiatus, it has been a great joy not only to rediscover the familiar people and places of my childhood but also to discover many new ones at MIT. I am deeply indebted to the friends and mentors who have made these first two years of graduate school so rewarding and helped me along the way to finishing this thesis. First and foremost, I am grateful to David Clark for his invaluable guidance, brilliant insights, and unfailing patience and good humor. I could not have asked for a better advisor. I am also thankful for the counsel and friendship of the other members of the Advanced Network Architecture group, especially Jesse Sowell, Shirley Hung, and Susan Perez. The students, faculty, and administrators of the Technology and Policy Program have been a constant source of academic and social support, and I feel incredibly lucky to have been part of such a fun, flexible, and intellectually stimulating program. Finally, I acknowledge the financial support of the Office of Naval Research under award number N00014-09-1-0597.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

> I can imagine no reason why an anonymous leaflet is any more honorable, as a general matter, than an anonymous phone call or an anonymous letter. It facilitates wrong by eliminating accountability, which is ordinarily the very purpose of the anonymity.
>
> —Justice Antonin Scalia, *McIntyre v. Ohio Elections Commission*, 1995

In Peter Steiner's iconic 1993 *New Yorker* cartoon, a black dog, sitting at a computer, instructs a fellow canine: "On the Internet, nobody knows you're a dog." In the nearly twenty years since the cartoon's publication, the Internet has grown at an astonishing rate, welcoming millions of new users and thousands of new applications, but its fundamental architecture remains largely unchanged, including those elements that enable the online anonymity highlighted by Steiner's dogs. It is still possible to browse and post information on the Internet without anybody knowing who you are—a feature that can prove deeply liberating in some circumstances, but highly problematic in others. In a speech on Internet freedom, United States Secretary of State Hillary Clinton (2010) noted, "On the one hand, anonymity protects the exploitation of children. And on the other hand, anonymity protects the free expression of opposition to repressive governments. Anonymity allows the theft of intellectual property, but anonymity also permits people to come together in settings that give them some basis for free expression without identifying themselves."

Online anonymity can be hugely beneficial at times but, as Clinton points out, it can also pose serious security threats. As the number of Internet users and applications has increased rapidly, so too, have the number of instances of online malware, denial-of-service attacks, espionage, spam, harassment and bullying. It can be more difficult to hold online users accountable for these types of misbehavior when they are anonymous. Terminating their online identities—which may take the form of anything from e-mail addresses to Facebook accounts—is often an extremely ineffective means of punishment, since malicious users can usually create new identities instantly, at no cost, and immediately resume their previous behavior. In light of the challenges of holding online users accountable for their actions, some advocates have called for the Internet's underlying protocols to be redesigned with accountability as a higher priority at the network layer.

1.1 Accountability at the Network Layer

Without effective mechanisms for holding anonymous Internet users accountable for their actions, some security proponents have proposed that the only solution to this accountability problem is to strengthen the attribution capabilities of the Internet's architecture, so that all online activity can be traced back to the responsible user's real identity. Former Director of National Intelligence for the United States Mike McConnell (2010) advocated such an approach in *The Washington Post*, writing: "We need to reengineer the Internet to make attribution, geolocation, intelligence analysis and impact assessment—who did it, from where, why and what was the result—more manageable." Researchers at Carnegie Mellon have proposed a design for a "Future Internet Based on Identification" based on this model, in which users are given Internet "ID cards" by the government that encode their real identities (He, Leon, Luan, & Sun, n.d.). Some countries have already taken steps to try to retrofit related user attribution schemes on top of the current network. South Korea has implemented its own attribution methods under the Law on Internet Address Management, which requires Korean websites that have more than 100,000 daily visitors to record users' real names and national identification numbers (Fish, 2009, p. 85). In 2009, several news sites in China instituted the same requirements with the announcement that they were acting under a "confidential directive issued . . . by the State Council Information Office" (Ansfield, 2009). It seems likely, therefore, that proposals like McConnell's to embed attribution mechanisms more uniformly and universally at the network layer to improve online accountability could garner support from many stakeholders in many other governments worldwide.

Knake (2010) likens such visions of "perfect attribution" to "the idea of giving packets license plates . . . Access to the network would require authentication, and each packet produced by the user would be traceable back to that user." Such an approach is technically feasible, Knake argues, but not desirable for two primary reasons: First, it would have tremendous implications for user privacy, turning the Internet into "the ultimate tool of state surveillance," and second, it would be unlikely to have any significant impact on our ability to identify criminals and attackers. "Ultimately, such a system would restrict the freedom and privacy of most users, while doing little to curb criminal elements or state actors who would find ways around the system," he concludes (Knake, 2010). Other researchers have raised similar concerns about the dangers and shortcomings of network-layer attribution schemes. "A public, personally identifiable packet-level mechanism is neither appropriate nor particularly needed or helpful," Clark and Landau (2010) argue, noting that such a system would do little to help with tracing multi-stage attacks, which go through multiple computers. Thus, embedding a strong attribution scheme at the network layer could fail to solve some of the most pressing security issues associated with cyber attacks, while simultaneously posing clear threats to the personal anonymity enabled by the current Internet architecture.

Both anonymity and accountability play important roles in sustaining the Internet's functionality. Without anonymity, a broad swath of users, ranging from activists living under oppressive regimes to people wishing to discuss their sensitive medical conditions, might be unable to pursue their online activities in comfort and privacy. However, this anonymity also encourages "anti-social" action, or malicious behavior, on the Internet (Tresca, 1998). Anonymity lowers users' inhibitions, both when it comes to expressing opinions, emotions or personality traits they might otherwise be embarrassed to display, but also when it comes to exhibiting hostile and damaging behaviors they might otherwise suppress for fear of repercussion (Suler, 2004). Without accountability, it is impossible to curb the numerous forms of online misconduct that interfere with users' online experiences and threaten the continued utility of the network. Reconfiguring the Internet to eliminate anonymous activity would perhaps help mitigate some of these negative behaviors, but it would likely do so at the expense of many of the important, positive anonymous behaviors enabled by the Internet. Clark and Landau (2011) point out that even if a network-layer attribution scheme is designed with the best of intentions, reducing cyber crime and deterring attacks, for instance, "once a mechanism for attribution is put in place, we must expect that it will be used differently in different jurisdictions." Clinton alluded to a similar risk in her 2010 speech, stating that, "Those who use the Internet to recruit terrorists or distribute stolen intellectual property cannot divorce their online actions from their real world identities. But these challenges must not become an excuse for governments to systematically violate the rights and privacy of those who use the Internet for peaceful political purposes."

How can we promote accountability on the Internet to more effectively prevent continued misbehavior, without sacrificing all the benefits afforded by online anonymity? One of the central arguments of this thesis is that achieving this goal requires us to implement a variety of different, context-specific accountability mechanisms at the Internet's application layer, rather than a single, uniform mechanism at the network layer.

1.2 Accountability at the Application Layer

Clinton's recognition that there are circumstances where strong authentication and rigorous attribution schemes are needed and other cases where these same mechanisms would be both harmful and inappropriate is echoed in the United States' 2011 "National Strategy for Trusted Identities in Cyberspace" (or "NSTIC"). The NSTIC lays out a proposed framework for an "Identity Ecosystem" for the Internet, but does not recommend a single, centralized authority for authentication of real-world identities, or a requirement that users identify themselves with a nationally recognized credential, in the manner of South Korea or China. Instead, the NSTIC recognizes that there is a great diversity of Internet activity which merits a diversity of identity schemes. It states:

> There are many Internet transactions for which identification and authentication is not needed, or the information needed is limited. It is vital to maintain the capacity for anonymity and pseudonymity in Internet transactions in order to enhance individuals' privacy and otherwise support civil liberties. Nonetheless, individuals and businesses need to be able to

> check each other's identity for certain types of sensitive transactions, such as online banking or accessing electronic health records. (2011)

It is this variety of Internet transactions, and the desire to tailor identity schemes with the appropriate degree of both anonymity and accountability for each of these transactions, that motivates our approach to implementing accountability mechanisms at the application layer of the Internet, rather than the network layer.

Affixing "license plates" to packets or otherwise embedding accountability at the network's lower-layer architecture imposes a single, uniform identity mechanism across the entire Internet, meaning that every online user, website, application, transaction, and communication is subject to the identical degree of accountability. This one-size-fits-all approach is at odds with the rich diversity of Internet applications and their vast range of varied functions. Clark and Landau (2011) explain:

> Some applications such as banking require robust mutual identity. Other sites need robust identity, but rely on third parties to do the vetting, e.g., credit card companies do so for online merchants. Some sites, such as those that offer information on illness and medical options, are at pains not to gather identifying information, because they believe that offering their users private and anonymous access will encourage them to make frank enquiries.

Forcing the same identity framework on medical advice websites or political discussion forums that we do on online banking sites fails to allow for these differences of function or account for the fact that these applications necessitate different degrees of anonymity and accountability. However, when online identity schemes are tailored individually for different applications, it becomes possible to design systems with the appropriate balance of anonymous protections, real-identity attribution capabilities, and corresponding accountability mechanisms. It becomes possible to treat banking applications, political forums, and medical support websites differently. It becomes possible to share the burden of holding users accountable across a number of different points of control in the network. Perhaps most notably, it becomes possible to create identity schemes that allow for the coexistence of both accountability mechanisms and anonymity protections, instead of sacrificing all anonymity in pursuit of perfect attribution.

The notion of creating identity systems for individual Internet applications is not a new one—in fact, it is the basis for many online applications today. A user who wants to join the popular social networking application Facebook first has to create a Facebook profile, an online identity specific to that application. Similarly, Internet users who wish to send and receive e-mail must first create an application-specific identities in the form of e-mail addresses. Indeed, most websites you visit or applications you use probably require or encourage you to create a new identity, or account. So, if there is already a system of application-layer identity schemes—if it's already possible for individual applications to tailor identities appropriate to their function—then why does accountability remain such a problem on the Internet? The answer to this question lies primarily in the widespread misconception that it is impossible to create online identities that are both anonymous and accountable. By examining the strengths and weaknesses of existing application identity schemes, it is possible to identify effective methods of establishing anonymous-accountable online identities that leverage the power of multiple control points on the Internet and can be used to improve the identity mechanisms of the vast number of applications which benefit from both some degree of anonymity and some means of holding users accountable.

1.3 Thesis Organization

This thesis begins by laying out a four-quadrant framework for understanding the interplay between accountability and anonymity of online identities; it then reviews a series of case studies of accountability mechanisms used by popular Internet applications, and concludes with a discussion of different application design patterns for accountable identities derived from these case studies.

The second chapter addresses the notion of an accountability-anonymity tradeoff and proposes an alternative, four-quadrant model of accountability-anonymity axes intended to more accurately capture the nuances and diversity of online identity schemes; it also looks at different control points in the application layer and addresses the problem of "discardable" identities in cyberspace and the ease with which bad actors can create new, free identities as soon as their original ones are terminated.

The third, fourth, fifth, and sixth chapters present case studies of four popular applications: e-mail, the virtual community Second Life, Facebook, and review site Yelp, respectively. These case studies review different types of malicious activity observed in each of these applications and then examine the identity schemes and associated accountability mechanisms employed by each to help mitigate or deter this misbehavior.

The final chapters explore broader methods of combining anonymity and accountability drawn from these case studies. The seventh chapter discusses potential design patterns for applications that allow users to decide on a personal trade-off between how much they invest in a given online identity and what privileges are associated with that identity, so that larger investments lead to greater privileges within the context of a given application, while users who make smaller investments are given correspondingly fewer capabilities. The eighth chapter focuses on conditional anonymity schemes, including methods of identity escrow and cryptographic protection of real identities. The ninth chapter addresses issues of accountability in identity management schemes and aggregated online identities. Finally, the tenth chapter summarizes the key lessons gleaned about how we can best characterize the space between perfect accountability and complete anonymity for Internet identity schemes, as well as how these schemes can be implemented—and by whom—to provide different combinations and forms of anonymity and accountability suitable to various online contexts and applications.

Chapter 2

The Anonymity-Accountability Axes

> The problem isn't anonymity; it's accountability. If someone isn't accountable, then knowing his name doesn't help. If you have someone who is completely anonymous, yet just as completely accountable, then—heck, just call him Fred.
>
> —Bruce Schneier, "Anonymity Won't Kill the Internet"
> *Wired*, Jan. 12, 2006

There is a common perception that online anonymity protections are irreconcilable with effective accountability mechanisms, derived largely from the broader idea of a "privacy-security" trade-off. Former National Security Agency code-breaker Ed Giorgio, who worked closely with McConnell on cyber strategy, described this trade-off succinctly to a reporter from *The New Yorker*, explaining: "We have a saying in this business: Privacy and security are a zero-sum game" (Wright, 2008). This claim that security and privacy stand in direct opposition to each other persists to an almost surprising extent, given how often it has been called into question and criticized by experts. Schneier (2008) writes,"Security and privacy are not opposite ends of a seesaw; you don't have to accept less of one to get more of the other . . . Security affects privacy only when it's based on identity." It is in the online identity space, therefore, that this trade-off between security interests, as represented by accountability and attribution advocates, and privacy rights, as championed by anonymity supporters, remains starkest and least reconcilable.

Davenport (2002) embraces the idea that anonymity and accountability constitute a zero-sum direct trade-off. He argues:

> By allowing anonymous communication we actually risk an incremental breakdown of the fabric of our society. The price of our freedoms is not, I believe, anonymity, but accountability . . . Accountability requires those responsible for any misconduct be identified and brought to justice. However, if people remain anonymous, by definition, they cannot be identified, making it impossible to hold them accountable.

2.1 The Four-Quadrant Framework

The widespread belief in this direct trade-off and the corresponding notion that it is necessary always to choose either accountability or anonymity in designing an Internet identity scheme may be partially responsible for the failure of many Internet applications to incorporate both features. However, the idea that accountability and anonymity are a zero-sum game has also been criticized. Farkas, Ziegler, Meretei, and Lörincz (2002) argue that "full anonymity may present a security risk that is unacceptable in certain applications; therefore, anonymity and accountability are both needed." This sentiment is echoed by Johnson, Crawford, and Palfrey (2004), who write:

> Anonymity does not need to be prohibited to allow accountability. There is no particular reason why a receiver needs a real-world identification of the source in order to make decisions about whether to accept a message or not. We see a key difference between authentication, on the one hand, and identification, on the other. All we need is to be able to tell the difference between the case in which a speaker (message sender) stands by a verifiable reputation (including the reputation of a pseudonym) and the case in which there is no way to tell anything with confidence about the source of the communication.

This insight that accountability can exist even in the absence of strong authentication is an essential one for designing identity schemes and associated accountability mechanisms that are appropriate for many of the applications in existence today. The pseudonymous reputation systems that Johnson et al. emphasize are one example of how this can be achieved in online applications. For instance, popular review site Yelp supplements every review with a detailed description of the "reputation" of the user who posted it, including information ranging from how long they've been a Yelp member and how many reviews they've posted, to the distribution of different rankings in their reviews. Notably, none of this reputational information is necessarily tied to the user's real-world identity.

Reputation systems are not the only means of combining accountability and anonymity in online identity schemes, however, and several other applications have found equally clever and effective methods for achieving their own balance. To clarify the range and variety of accountability and anonymity options available to application designers, we propose an alternate framework to the "zero-sum" model that conceives of anonymity and accountability as a direct trade-off, wherein having more of one necessitates having less of the other. This traditional, "one-dimensional" notion of accountability and anonymity is illustrated in Figure 2-1.

An alternative, two-dimensional framework for understanding the more complex ways in which these features can be combined instead of being treated as a direct trade-off, is illustrated in Figure 2-2. The accountability-anonymity axes represent spectrums along which different degrees of anonymity and accountability may be combined with each other, and the resulting four quadrants provide a useful framework for classifying and analyzing different online identity schemes. Most interestingly, where

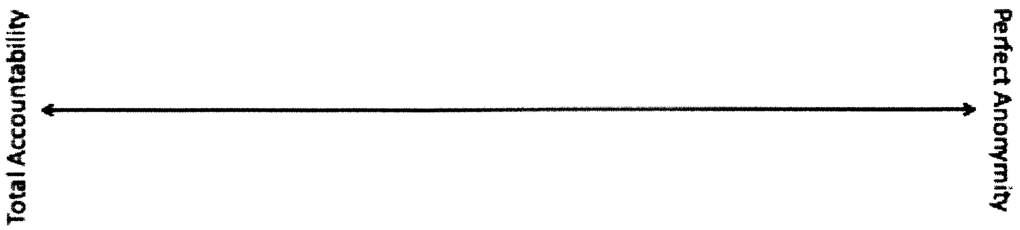


Figure 2-1: Traditional framework for thinking about accountability and anonymity in cyberspace as zero-sum game.

the "traditional" zero-sum framing allows only for identities that fall into either the upper left (strong accountability-weak anonymity) or lower right (strong anonymity-weak accountability) quadrants, this framework opens up two additional quadrants, providing a richer and more nuanced perspective on the interplay between accountability and anonymity. These new quadrants, especially the upper right one which combines both strong accountability mechanisms and strong anonymity protections, are essential for understanding both the challenges and the full range of possibilities permitted by an application-layer approach to accountability. To more clearly illustrate this range, it is helpful to populate the accountability-anonymity axes with some representative applications in each of the quadrants, as shown in Figure 2-3.

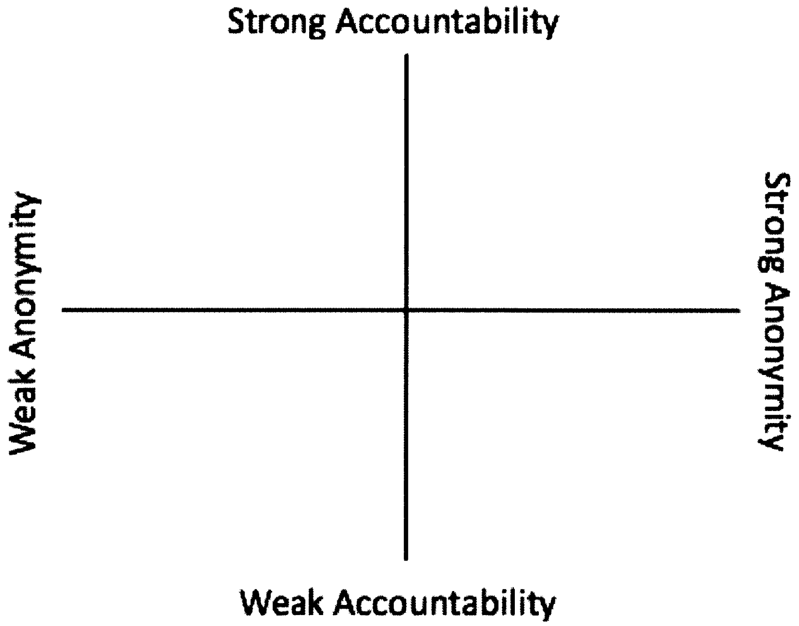


Figure 2-2: Proposed alternate framing of four-quadrant space for online accountability and anonymity.

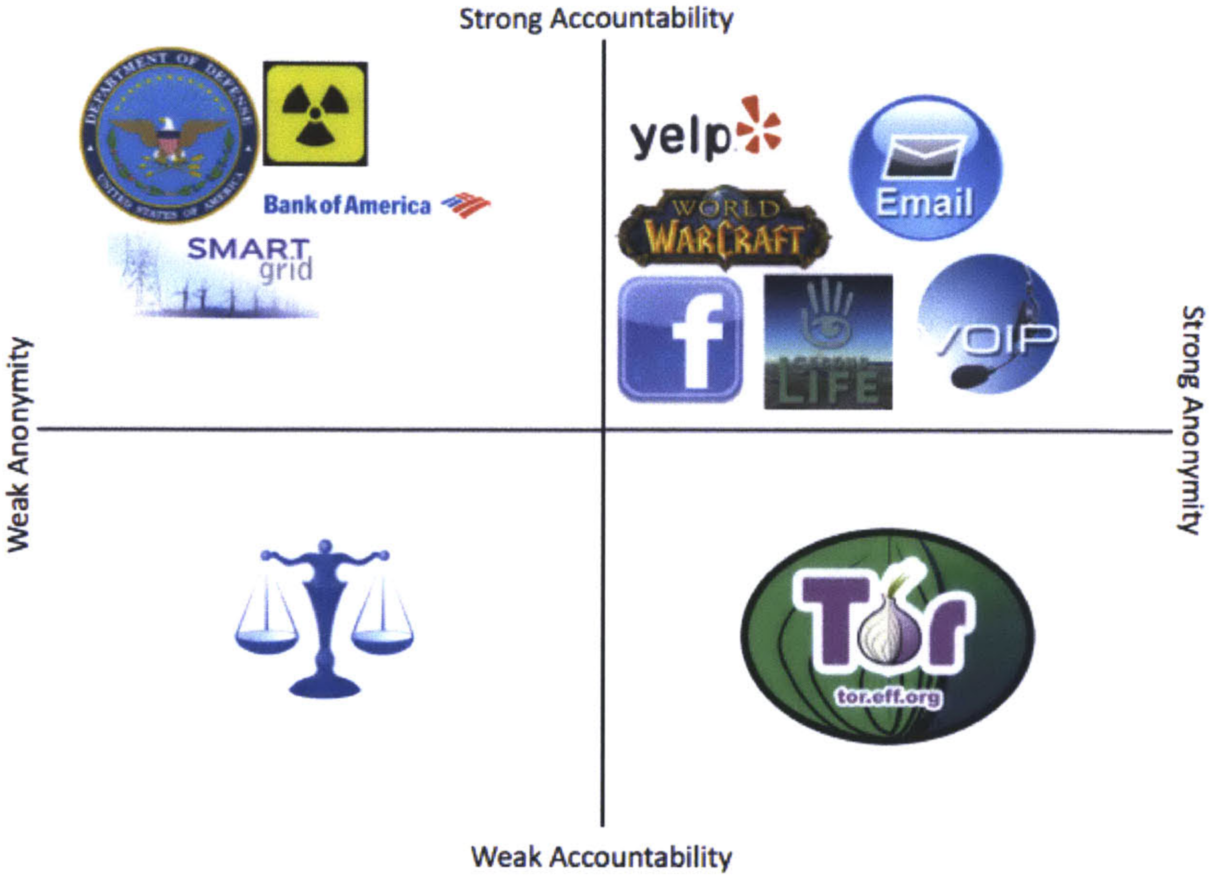


Figure 2-3: The anonymity-accountability quadrants populated with representative applications.

Identifying example applications that fall within each quadrant emphasizes the role of the upper left and lower right quadrants in allowing for applications that do, in fact, fall at the far ends of the traditional zero-sum framing of accountability and anonymity. For instance, the four-quadrant model still allows for online identities that do not exhibit anonymity protections and instead require strong accountability in the form of close ties to a real-world identity and robust authentication. A cybersecurity report published by the Center for Strategic and International Studies advocates "making authentication requirements proportional to risk," so that "high-risk situations require strong authentication, while the lowest-risk situations require no authentication" (Langevin, McCaul, Charney, Raduege, & Lewis, 2008, p. 64). The top left quadrant, then, is home to the applications that would pose the greatest security risks if accessed by unauthorized or malicious users. These might include military and classified data networks, banking applications, nuclear power plant systems, and electric grid networks, as well as other critical infrastructure elements. In cases like these, where all legitimate parties involved in the online transaction—for instance, both a bank and a bank account holder—would reasonably wish for strong and secure authentication of the other participating parties, anonymity is not called for. Instead, robust authentication schemes allow for stronger security and accountability mechanisms that depend largely on identifying the responsible actors in the real world and holding them accountable for their actions in a traditional legal and regulatory manner. The risks associated with unauthorized infiltration of these sorts of networks are simply too high, and the benefits of allowing anonymous activity too minimal, to merit implementing any weaker forms of online identification or alternative, pseudonymous accountability mechanisms.

By contrast, the lower right quadrant is populated with applications where anonymity trumps accountability and users may enjoy very strong identity protections while encountering relatively weak accountability mechanisms. Few applications are as synonymous with strong online anonymity protections as The Onion Router (or Tor) software that routes users' Internet packets through a global network of volunteer servers to conceal their originating location. Tor's operators describe the application on its website as a service that "protects you by bouncing your communications around a distributed network of relays run by volunteers all around the world: it prevents somebody watching your Internet connection from learning what sites you visit, and it prevents the sites you visit from learning your physical location." In other words, the express purpose of Tor is to protect the anonymity of its users and, in doing so, it has played an important role in empowering activists and dissidents in oppressive or autocratic regimes to speak freely and organize online, especially in countries such as Iran and Egypt. Tor developer Jacob Appelbaum explained in an interview:

> Because Twitter and other websites were blocked, people in Egypt actually used Tor as a proxy for their web browser. They knew that they could install Tor and they would be able to get past the Internet censorship in their country, which was their primary concern. ... Tor is only as secure as the protocols you send across it, except in certain cases such as this one where you know that the problem is directly between you and the Internet. In that case, Tor is extremely secure and no matter what you are doing over Tor you are almost certainly better off than the government that might arrest you for behavior you do on the Internet or for other things that they would be able to detect and log and then later analyze. (Zahorsky, 2011)

Just as the examples of banking and military applications illustrated the value of strong authentication in some cases, the use of Tor by Egyptian dissidents demonstrates the importance of powerful anonymity protections in others. That is not to say there aren't real risks posed by the anonymity afforded to users by applications like Tor. "Anonymity is like a rare earth metal. These elements are a necessary ingredient in keeping a cell alive, but the amount needed is a mere hard-to-measure trace. In larger does these heavy metals are some of the most toxic substances known to a life. They kill," writes Kelly (2006). He continues: "There's a dangerous idea circulating that the option of anonymity should always be at hand, and that it is a noble antidote to technologies of control ... in every system that I have seen where anonymity becomes common, the system fails." It is worth looking more closely at Tor to understand how it has managed to escape this fate, even as its entire design is predicated on providing total anonymity for all of its users. After all, that same protection which allows dissidents to revolt against Hosni Mubarak can also potentially enable users to browse for illegal materials or engage in damaging and dangerous online attacks. However, while Tor is undoubtedly used for illegal and illicit purposes, its design minimizes the risks associated with users' actions based on one simple trait: Tor is slow (Dingledine & Murdoch, 2009). Slowness is usually considered a design flaw of Internet applications, not a positive feature, but Tor's slow speeds mean that it is nearly impossible for malicious actors to use its anonymity protections to inflict any high-traffic damage, such as launching a large-scale botnet or denial-of-service attack. In other words, Tor's slowness reduces many of the risks associated with its powerful anonymity protection, effectively eliminating some of its most dangerous possible uses and thereby allowing it to thrive even in the absence of rigorous accountability mechanisms.

The drawbacks to the weak accountability mechanisms of applications in the lower right quadrant, like Tor, are mitigated to a certain extent by the positive anonymity protections they provide. However, in the lower left quadrant, where both accountability and anonymity are weak, there are no mitigating factors. This quadrant is the domain of Internet jurisdiction disputes, when malicious actors can be clearly identified yet it is still impossible—or extremely difficult—to hold them accountable for their online actions because they are located outside the aggrieved party's jurisdictional borders. These jurisdictional disputes are often a primary reason for inadequate Internet accountability mechanisms. For instance, people who rent out botnets to send spam must provide some form of identification in order to receive payment for their services and this identification should, in turn, make it possible to perform some degree of attribution and to hold the responsible actors accountable. For this reason, Clark and Landau (2011) explain that "Spammers' protection comes not from anonymity, but from jurisdictional distance or legal ambiguity ... Even if we were to push for a variant of the Internet that demanded very robust identity credentials to use the network, tracing would remain subject to barriers that would arise from variation in jurisdictions." The jurisdictional issues that inhibit the online accountability efforts of applications that reside in the lower left quadrant are central to any discussion of Internet accountability schemes, though they are not the particular focus of this thesis. While they warrant greater discussion and further research in their own right, these ongoing jurisdictional problems also help highlight the crucial role of the anonymous-accountable identity schemes that reside in the diagonally opposite, upper right quadrant and, in general, do not rely predominantly on state intervention and legal regulations for implementing accountability.

It is this fourth quadrant, in the upper right corner of the axes, which combines elements of strong anonymity protections with strong accountability mechanisms and is the focus of this analysis. Perhaps the first thing worth noting about this quadrant—though it may already seem redundant at this point—is simply that it exists at all, that there are in fact means of creating online identities that afford users both anonymity and accountability. Such identities are not just a hypothetical possibility, they exist today, in some of our most popular online applications, including Facebook and Yelp, and the subsequent case studies are largely devoted to understanding how such anonymous-accountable identities work in current online applications, as well as how they could be improved. Given that such identities schemes are already so widespread, it may seem unnecessary to devote time to emphasizing their existence, but even though so many of our online identities fall within the realm of the upper right, accountable-anonymous quadrant, pseudonymous and anonymous online identities are often dismissed as entirely unaccountable. Even executives from major Internet application companies such as Facebook and Google have defended their respective real-name policies—in which users are required to associate their online identities with their legal names—as necessary to promote stronger accountability, the underlying assumption being that there is no other means of holding users accountable beyond stripping them of their anonymity (Pfanner, 2011).

To support the assertion that it is possible to combine both accountability and anonymity in online identities at the application layer, we investigate several case studies of popular Internet applications and the ways in which they have successfully—or unsuccessfully—implemented anonymous-accountable identity schemes. From these case studies, we will then extract some more general methods and strategies for designing these types of online identities for different sorts of Internet applications. In this quadrant, it is important to note that, to an even greater degree than in either of the neighboring two quadrants, it is essentially impossible to make any claims of achieving perfect or absolute forms of either anonymity or accountability. This is why the axes are labeled as going towards "strong" accountability and anonymity rather than "complete" or "total" instances of either quality. However, even though the applications housed in this quadrant may not feature perfect accountability or anonymity, it is possible to adjust the relative strength of each, as well as the manner in which each is implemented, to suit the functionality and design of a variety of different types of Internet applications.

Importantly, the four-quadrant framework for thinking about online identities laid out in this section is only applicable to an approach to accountability that occurs at the application layer. The ability to populate these four quadrants and achieve different balances of accountability and anonymity for different applications depends entirely on implementation at the application layer; accountability mechanisms that reside at lower levels of the network and thereby institute a one-size-fits-all solution across the entire Internet do not allow for this variety of different, customized approaches.

2.2 Application-Layer Points of Control

Implicit in the notion of accountability is the involvement of at least two parties: one that is being held accountable and another that is holding them accountable. In the physical world, we are held accountable for our actions by a variety of different actors; governments and courts of law hold us accountable for illegal and criminal actions, friends and family hold us accountable for violations of social norms, colleagues hold us accountable for workplace missteps, and schools and universities hold us accountable for academic failings. In other words, there are many different types of behaviors that, for various reasons, we are discouraged from engaging in, and we are held accountable for these behaviors by a variety of different institutions, individuals, and groups of people. Similarly, there are a variety of institutions, individuals, and groups of people that can hold online identities accountable for the Internet actions associated with them. In this section, we will examine some of the different points of control at the application layer that can exert their various powers to hold online actors accountable in different ways.

Scholars have identified two primary dimensions of accountability: answerability and enforceability. Answerability refers to actors' obligation to account for their actions, while enforceability encompasses the mechanisms used to punish or sanction those actors for harmful or illegal actions (Newell, 2006). Of these, answerability is the trickier one to translate into cyberspace. In the context of political accountability, it refers generally to transparency, or requiring political actors to describe and justify their actions to their constituents. Online, this sort of transparency of actors is both easier and harder to achieve: easier in the sense that users, in many cases, cannot effectively conceal their actions or decouple them from their online identities, but also much harder in the sense that, as discussed earlier, attributing these actions to real-world identities can be extremely difficult. For instance, if we consider trying to hold spammers accountable, it is relatively easy to identify the "online identity" responsible for the spam (that is, the "from" e-mail address), but often much more challenging to pinpoint the person (or people) in the real world who should have to answer for this action. It may be impossible, in some cases, to trace online activity back to the specific responsible person, but it is almost always possible to trace that activity back to something, be it a machine, an intermediary party, or a virtual identity. In other words, there are always entities answerable for every online action, even if these entities are not necessarily the direct perpetrators. This interpretation of answerability for the Internet informs the range of enforceability mechanisms possible online, both by suggesting that enforcement centered on holding responsible actors accountable may be of limited value and by revealing a substantial set of alternative targets for enforcement mechanisms—those machines, intermediaries, and virtual identities that are potentially answerable for online activity.

To understand how accountability can be enforced in the context of online applications it is first necessary to identify which actors will carry out this enforcement. Each of these actors serves as a control point, exerting some type of power or control over the application which can be brought to bear for the purposes of punishing and sanctioning users. Each control point is capable of enforcing some degree of either vertical or horizontal accountability, where vertical accountability describes "a relationship between unequals," that is a more powerful actor holding some less powerful actor accountable (or vice versa), and horizontal accountability refers to "somebody holding someone else of roughly equal power accountable" (Schedler, 1999, p. 23).

The most straightforward sources of vertical accountability for an Internet application are its designers, owners, and operators—the actors who write and update an application's architecture, or code. In some cases, these roles may all refer to a single entity. For instance, Facebook, Inc. serves as designer, owner, and operator of the social networking site Facebook. Other applications, notably e-mail, were designed with an open set of standards and are therefore "owned" by no one and operated by a decentralized network of individual users and server administrators. For this reason, it can be useful to distinguish between an application's designer and its operator, even though these roles are often conflated in a single control point. Application designers play a crucial role in dictating the initial accountability mechanisms and identity schemes associated with a given application. Through their choices, designers embed in the very code and protocols of an application how much and what kind of information a user will be required to provide in order to create an identity, what capabilities and privileges that identity will have, what tools other users will have to keep those capabilities and privileges in check and respond to misbehavior, and what information about users, their online identities, and their actions will be stored privately or made public to other users. In determining and embedding these initial requirements and specifications, application designers exert tremendous influence in forming the norms, behaviors, and expectations that come to govern interactions in their applications. While their code and design choices can be adjusted or revised later on, these norms and expectations may be more difficult to alter once users have already adopted them. Of course, it is nearly impossible for application designers to anticipate, prior to deployment, all possible modes of misbehavior that may crop up or evolve over time among their applications' users—this is where application owners and operators step in.

Application owner and operators are responsible for the continuously revising and updating an application's code and Terms of Service, as well as responding to user concerns and complaints. Many Internet applications are owned and operated by private companies; for instance, video game developer Blizzard Entertainment maintains the popular role-playing application World of Warcraft, Linden Lab similarly operates the virtual world Second Life, and the non-profit organization WikiMedia Foundation runs the collaborative user-written encyclopedia Wikipedia. These entities exercise many of the same powers as the application designers—the ability to alter an application's identity scheme and the privileges afforded to its users, for instance—but lack the designers' initial, formative impact on user norms and expectations. In many cases, however, organizations do design the applications they subsequently own and operate. Additionally, owner-operators often store logs of user activity and associated information, such as originating IP addresses for user activity, which can serve as valuable sources of data about the applications' users and also, at times, aid attribution efforts or serve as legal evidence in court. Operators of many applications also respond to complaints or "abuse reports" from users and must determine which of these concerns merit further attention or disciplinary measures. User input may also, at times, motivate operators to tailor their application's design to user requests and proposals. Finally, these operators can exercise the ultimate online punishment of terminating misbehaving online identities, or accounts, belonging to users who misbehave egregiously or repeatedly within their application, though this is not always an effective sanction, particularly when users can easily and cheaply create new identities.

Designers and operators serve as the primary sources of vertical accountability for online applications; these control points clearly have more power than individual end-users when it comes to defining an application's capabilities and associated identities. However, many Internet applications also benefit from a strong culture of horizontal accountability, in which users are partially held accountable by actors who are their equals in power—their fellow users. Individual end-users represent a remarkably powerful control point in several online applications, especially those which do not have strong centralized owner-operator control points, such as e-mail, and those which rely on primarily bilateral user interactions, where individuals can often regulate whom they do or don't want to interact with to a great extent. Johnson et al. (2004) describe this phenomenon as "peer production of governance" and predict that the "aggregation of numerous individual decisions about who to trust and who to avoid will create a diverse set of rules that most accurately and fairly serves the interests of those who use the online world." By blocking unwanted interactions, indicating to others when users are misbehaving, and constructing powerful social and cultural norms, end-users can hold each other accountable in ways that support and extend the accountability mechanisms enacted by designers and operators. However, it is worth noting that these end-user capabilities are generally only made possible by technical tools provided by the designers and operators, such as "block" commands to avoid contact with specific other users or embedded reputation systems. Newell (2006), describing political accountability, wrote: "To be effective, horizontal accountability needs to be buttressed by strong vertical accountability." The same holds true for Internet accountability: end-user efforts at horizontal accountability are most effective when they are enabled and reinforced by an application's design and ownership.

Besides the centralized, top-down designer and operator authority and the decentralized, bottom-up end-user governance, Internet applications are influenced by several other control points which can also play important roles in holding online actors accountable. Government actors and courts can serve as a powerful control point by enacting and enforcing legal regulations such as the United States' CAN-SPAM Act and state cyberbullying laws, the European Union's e-commerce directive, and Pakistan's Prevention of Electronic Crimes Ordinance. These law-based enforceability mechanisms are limited both by jurisdictional scope and attribution capabilities but, nonetheless, they have been effectively leveraged in some cases to hold Internet users accountable for actions ranging from spamming to cyberbullying, online defamation, and trolling. A range of other intermediary control points can also help give rise to greater accountability within certain types of applications. For instance, some community-based applications develop participatory, emergent governance structures, such as the system of user moderators and bureaucrats in Wikipedia and the user-organized Second Life Exchange Commission. Additionally, intermediate control points such as e-mail server administrators and Internet Service Providers can play an important role in monitoring and sanctioning some forms of user misbehavior.

Some of these control points are universally applicable—every application has a designer (or designers) and users—while others are much more application-specific—many applications, for instance, do not feature independent server administrators or internal exchange commissions. In other words, different applications have different points of control, each of which may wield different kinds of power. An advantage of customized, application-layer accountability mechanisms is that they can be tailored to each application's individual set of control points and designed in such a way that all of these actors support and reinforce each others' efforts to hold users accountable. In this manner, it is possible to leverage and combine the strengths of several different control points to compensate for the limitations and weaknesses of each. For instance, application operators are often capable of enforcing relatively powerful punitive measures, ranging from removing public content to curtailing users' privileges to account suspension or termination. These operators are often limited, however, by how much time and human capital they can devote to monitoring the actions and behavior of individual users. Particularly for the most popular applications, whose users number in the hundreds of millions, this can be an impossibly burdensome task for a single operating company. For instance, more than 250 million photos are uploaded to Facebook every day. Even if every single one of the company's roughly 2,000 employees devoted all of their time to reviewing each of these photos, they would probably still be unable to reliably identify and delete all inappropriate or illegal content on the site.

By contrast, the control point constituted by individual Facebook users has nearly opposite strengths and weaknesses: the application's end-users have relatively little punitive power—they cannot remove offending photos posted by other users or shut down other people's accounts—but there are more than 800 million of them and, collectively, they spend a staggering amount of time on the site, viewing posted content and observing the behavior of other users. Thus, Facebook can, in some sense, outsource content monitoring to its users, allowing them to report particularly offensive or problematic behavior and content to the company, which can then decide what punitive measures, if any, are appropriate. Similar "notice-and-take-down" systems are implemented by many other popular applications, including video website YouTube, to combine the diffuse monitoring capabilities of end users with the centralized enforcement capabilities of the application operators. Approaches like these, that leverage the power of multiple control points, can be especially useful when trying to create a variety of effective, customized accountability mechanisms for different types of Internet applications.

2.3 The Problem of Discardable Identities

Few virtual identities have garnered as much media attention and notoriety as Mr. Bungle, the avatar who perpetrated the much-publicized 1993 cyber rape in the text-based online community LambdaMOO. Mr. Bungle used a "voodoo doll" subprogram that allowed him to control the actions of other users in the community besides himself, to narrate a lengthy and graphic series of "forced" sexual encounters involving the other avatars. These encounters occurred only in the words typed by Mr. Bungle that appeared on each user's computer screen, but they led to tremendous outcry among the community's users and Mr. Bungle's account was deleted shortly thereafter by one of LambdaMOO's moderators. A few days later, a new avatar named Dr. Jest appeared in LambdaMOO. "There was a forceful eccentricity to the newcomer's manner, but the oddest thing about his style was its striking yet unnameable familiarity," Dibbell (1993) writes of Dr. Jest, adding that, "when he developed the annoying habit of stuffing fellow players into a jar containing a tiny simulacrum of a certain deceased rapist, the source of this familiarity became obvious: Mr. Bungle had risen from the grave." LambdaMOO suffered from an accountability problem common to many Internet applications: discardable identities. After Mr. Bungle's account was deleted, Dibbell (1993) points out that all the user had to do was "go to the minor hassle of acquiring a new Internet account, and LambdaMOO's character registration program would then simply treat the known felon as an entirely new and innocent person."

If, as in LambdaMOO, an operator's ultimate means of redress is deleting a user's account or virtual identity for a given online application, there must be some checks on the user's ability to immediately and freely create a new identity for this to be an effective accountability mechanism, or form of punishment. In other words, creating a new virtual identity must have some cost to the user, otherwise users will essentially be able to behave as they please when using online applications with little or no fear of the consequences. As Landwehr (2009) points out, "To be accountable, the individual, company, or system component must pledge something of value—money, reputation, friendship—that can be forfeited in case of improper actions." For most online identities there are two possible different forms this cost can take: an investment of either money or time. In some cases, for instance with many social network sites, the time investments required to construct these identities may also be closely linked to users' real-life friendships, adding another element of value that users may risk by misbehaving.

Perhaps the most straightforward means of attaching a cost to virtual identities, and thereby making them less discardable, is simply to attach a monetary price to them. If creating an e-mail account cost $10—or even $1—most spammers would be more reluctant to risk termination of their e-mail addresses by sending spam from them. Similarly, requiring people to pay money for Wikipedia or Facebook accounts might reduce the amount of misbehavior, false postings, and harassment in these applications. However, it would also most likely cut down on the amount of legitimate, appropriate activity within these applications and drastically shrink their consumer base. Anderson (2008) explains, "Give a product away and it can go viral. Charge a single cent for it and you're in an entirely different business, one of clawing and scratching for every customer ... The moment a company's primary expenses become things based in silicon, free becomes not just an option but the inevitable destination." It is perhaps no surprise, then, that many Internet applications, in their eagerness to reel in large customer networks, are reluctant to impose financial costs on user identities to improve accountability. Paid identities often also have the effect of reducing anonymity, since credit card charges can be traced back to real-world entities. Furthermore, it is worth noting that charging fees for entry to online communities is not sufficient to eliminate online bad behavior. World of Warcraft players, for instance, are charged regular monthly fees for access to the game, but moderators continue to struggle with how to curb "griefing" behavior, in which players intentionally torment and sabotage other users for no reason other than causing trouble.

Another potential drawback to charging users for their identities is that it does not dissuade all users equally from engaging in destructive or discouraged behaviors. Instead, such a system has a stronger impact on poorer users and is much less effective at curbing the misbehavior of the rich. This is also true, to a certain extent, of identities that require investments of time—since richer users could presumably pay others to spend the necessary time to build new identities for them—but financial costs have the added disadvantage of completely eliminating the participation of any users who cannot afford to pay the identity fee. While both time and financial investment schemes potentially allow richer users more freedom to misbehave in cyberspace, time investment costs still permit poorer users to access Internet applications, so long as they behave appropriately within the context of a given application.

Requiring users to invest money in their online identities is not the only means of rendering those identities less discardable; investments of user time and energy in creating these identities and building up reputations for them can also be quite effective. One of the simplest ways of imposing this time investment is to require new users to wait through an "initiation period" of some set duration before allowing them full access to all the privileges of their accounts. The initiation period model does not prevent users whose accounts are deleted from forming new identities, waiting through the initiation period again, and then resuming their previous misdeeds. However, the cost of waiting through this initiation period before being able to engage in any further destructive behavior might, in some cases, be sufficient to deter further misbehavior, and at the very least, it could have the effect of slowing the rate of malicious activity. Users who are forced to invest the necessary time to fully activate their online identities are less likely to be willing to forfeit this investment simply to cause trouble and more likely to make some greater effort to protect their identities and shield them from termination.

Another time-investment approach involves collecting and publishing public reputation data about an application's account holders in order to allow other users to assess the value and trustworthiness of a given identity. Reputation systems operate on a similar principle to the initiation period model: the idea that users who have to invest a certain amount of time in building up a reputation for their online identities will be less likely to risk losing those identities by misbehaving. Instead of forcing users just to invest time in their identities by waiting out an initiation period, reputation systems require users to invest both time and energy in creating online reputations. For instance, review website Yelp displays relevant data about all of its users next to their reviews—data like how long they have been Yelp members, how many reviews they've written, and whether other users have found their reviews helpful. This system allows Yelpers to instantly assess the credibility of the reviews they read and also encourages users to build up credibility by investing more effort in writing reviews to develop their identities. While these sorts of reputation systems do not prevent users from creating new accounts simply to engage in malicious behavior—for instance, writing a glowing review for one's own restaurant—it is much easier for other users to identify these accounts and assess them accurately with access to detailed reputation data. Johnson et al. (2004) explain:

> As long as . . . individuals use systems that require those who interact with them to authenticate themselves and/or provide acceptable reputational credentials—using a contextually-appropriate mode of authentication—then everyone can decide when to trust someone (some source of messages) and when to filter someone else out of their online world altogether. Using such systems, we can collectively hold those with whom we interact online accountable for their antisocial actions (and for their failures to hold others accountable).

From an accountability standpoint, users who have invested the time and energy in building up strong reputations for their virtual identities will often want to preserve those reputations by protecting their online identities from deletion and avoiding negative attention from other users, so they are therefore more likely to be careful about not violating an application's terms of use or mistreating fellow users (Brennan & Pettit, 2004). Another feature of this model is that poorer users are equally able to build up and sustain positive reputations as richer users, although poorer users may still be less free to engage in misbehavior if rich users pay other people to build up the reputation of new accounts for them. However, unlike the fee model, poorer Internet users still have access to applications like Yelp that rely on time investment costs, and so long as they do not wish to behave in ways that result in their accounts being terminated, they are not necessarily at any great disadvantage when creating these identities.

One variation on imposing monetary or time costs on Internet identities involves imposing these costs on individual online actions, such as sending e-mail, rather than accounts. Instead of requiring users to invest time or money in creating their e-mail addresses, they could conceivably be asked to make some similar (though presumably smaller) investment in each e-mail they wish to send. A sender could either pay the equivalent of an e-mail postage stamp fee (financial investment) or instead be charged a "processing" fee (time investment) by performing some computational task that requires a certain number of CPU cycles per message (Dwork & Naor, 1993). This investment model is most applicable to behaviors like spamming that are problematic based on high frequency of a particular action, such as sending e-mail, rather than the action itself.

Combating discardability by creating costly virtual identities is a central goal of online accountability mechanisms and though each of these investment schemes leaves the door open for certain forms of misbehavior, especially on the part of richer users, imposing even small costs on online identities can play an important role in mitigating the amount of malicious online behavior. Publicizing users' investments in their online identities can also enable greater accountability while still permitting people to invest according to their individual abilities. For instance, an application might allow users to pay however much they wanted for an online identity on the condition that this information would be publicly available to all other users of the application. The virtual world Second Life has operated a similar system, permitting users to create either free or paid accounts but then posting whether or not an avatar had been paid for in its public profile. This approach combines elements of monetary investment and reputation systems and enables people to assess other users' investments and decide whether or not they wished to interact with or trust an account, based on how much has been invested in it. Publicizing the investments users make in their online identities by making available information such as the amount of time accounts have been active, the amounts paid for them, or the rankings assigned to them by other users, is a way for application designers to let investments serve as signals to other users of how discardable—and therefore how accountable—these identities are.

2.4 Accountability Beyond Attribution

Today's Internet is incapable of providing its users with either perfect accountability or complete anonymity—users always have some means of masking their identities but no reliable way to completely or permanently erase every trace of a packet's origin. Trying to re-engineer the network to enable more rigorous attribution of online activity would upset this balance and likely hinder the development of the increasingly rich diversity of Internet applications we currently enjoy. However, it is possible to better understand and strengthen the accountability mechanisms embedded within these applications without resorting to network-layer alterations. Close examination of several popular Internet applications and their associated user identity schemes reveals several ways in which accountability mechanisms can be coupled with significant anonymity protections without compromising the effectiveness of either, giving rise to the four-quadrant model of the anonymity-accountability axes. These axes provide a more nuanced and accurate characterization of the space between perfect accountability and complete anonymity for Internet identity schemes than the one-dimensional zero-sum model and provide a framework for further analysis of some online accountability mechanisms—such as identity investment-privilege trade-offs, discussed in chapter 7, and conditional anonymity schemes, described in chapter 8—that can be implemented without disregarding user anonymity.

The most important function of these axes is not to highlight one particular accountability mechanism as the "solution" for online applications but rather to emphasize the need for a diversity of such mechanisms to match the diversity of existing applications. The crucial job of application designers is determining what identity schemes will provide the appropriate combinations and modes of anonymity and accountability for their particular applications. By tailoring the accountability and anonymity of online identities to specific applications and their respective control points and users, it is possible to preserve the Internet's hallmark flexibility and versatility while still preventing troublemakers and criminals from overrunning the network. The problem is not that on the Internet, nobody knows you're a dog nor is it that packets don't have license plates. Accountability is a problem on the Internet because users too often do not make sufficient investments in their online identities to care about what happens to them. Ratcheting up those investments and rendering online identities less easily discardable is the central aim of mechanisms that fall within the upper right quadrant of the anonymity-accountability axes, and the contributions of this quadrant play a crucial role in enabling the effective implementation of accountability at the Internet's application layer.

Chapter 3

E-Mail

> ON 2 MAY 78 DIGITAL EQUIPMENT CORPORATION (DEC) SENT OUT AN ARPANET MESSAGE ADVERTISING THEIR NEW COMPUTER SYSTEMS. THIS WAS A FLAGRANT VIOLATION OF THE USE OF ARPANET AS THE NETWORK IS TO BE USED FOR OFFICIAL U.S. GOVERNMENT BUSINESS ONLY. APPROPRIATE ACTION IS BEING TAKEN TO PRECLUDE ITS OCCURRENCE AGAIN.
>
> —Maj. Raymond Czahor
> Chief, ARPANET Management Branch, DCA
> May 4, 1978

Unsolicited bulk e-mail, or spam, is one of the earliest forms of online misbehavior, dating back to the pre-Internet era of ARPANET. Then, it may have been considered worrisome because it did not pertain to the "official U.S. government business" the network was originally intended for but today, with an Internet whose functions have increased vastly beyond ARPANET's mandate, spam is a problem for different reasons. As the volume of e-mail spam has increased rapidly over the past three decades, three features of the unwanted messages have been identified as especially troublesome: their content, their consumption of Internet resources, and the threat they pose to Internet security. These concerns stem from the fact that spam messages sometimes include objectionable or pornographic content and they consume an enormous amount of network bandwidth, memory, and storage space, while often being used to transmit viruses and malware (Sorkin, 2001). Researchers have argued that spam can take a severe financial toll on its recipients, costing the global economy roughly $50 billion each year in "direct outlays, lost productivity, interruptions and wasted time" (Koomey, Alstyne, & Brynjolfsson, 2007). Some even suggest that the growing volume of spam could, at some point, become so problematic—cause such substantial economic losses, engender such strong user irritation, spread so many harmful viruses—that it would ultimately threaten the continued use of e-mail as a communication medium (Helman, 2009).

Considering how many problems spam poses and how long it has been around, it is not surprising that many companies and individuals have devoted—and continue to devote—considerable time and resources to fighting it. These efforts range from technical filters and blacklists to legislation and and, at times, they have seemed sufficiently promising as to engender considerable optimism from the anti-spam camp. At the 2004 World Economic Forum, Bill Gates announced: "Two years from now, spam will be solved" (Koomey et al., 2007). Eight years later spam has still not been solved; in fact, it has increased exponentially, to comprise roughly 80 percent of all e-mail traffic, totaling nearly 200 billion messages every day.

In response to the rapidly growing spam volume, industry actors and government officials across the world have developed a variety of technical and legal anti-spam measures, but in many cases these techniques have largely failed to curb the continued growth of the spam industry. Spammers have time and again shown themselves capable of outwitting and bypassing increasingly sophisticated technical solutions, ranging from regularly updated blacklists to advanced filters. Meanwhile, most legal solutions have gone largely unenforced. The failure of these numerous well-funded and well-researched attempts to reduce spam e-mail speaks to the difficulty of holding spammers accountable, a trait that is deeply entrenched in the underlying protocols and architecture of e-mail.

3.1 Why Does Spam Persist?

In retrospect, Gates' 2004 prediction may appear unduly optimistic, but surveying the vast body of research, growing industry efforts, and numerous legislative measures all dedicated to combating spam, it seems surprising that the anti-spam camp has not met with more success. The underlying reason spam has managed to persist in the face of so many well-resourced efforts to stamp it out lies in the architecture of e-mail, which has allowed spammers to transfer the bulk of the costs of sending spam onto mail recipients and administrators, as well as other intermediaries (Dickinson, 2004). At the most basic level, spam persists because it continues to be a profitable activity for spammers. For the most experienced and successful spammers, it can even be a very lucrative business: North Carolina-based spammer Jeremy Jaynes earned $750,000 per month, for a career total of roughly $24 million, prosecutors estimated at his 2004 trial (Helman, 2009). However, recent studies on the economics of spam suggest that Jaynes' case is the exception and not the rule; many spam campaigns seem to have relatively meager margins of profit and are therefore "economically susceptible to new defenses" (Kanich et al., 2008).

Three main factors contribute to making spam profitable for spammers. First, there is an exceedingly low marginal cost to the spammer of sending out additional e-mails—much lower than the cost of launching comparable unsolicited, bulk messaging campaigns by postal mail or telephone. Second, senders are not required to verify or authenticate their identity in order to send e-mails, and spammers are therefore very difficult to trace and identify. Third, e-mail allows for people to send e-mails to recipients whom they do not know, or rather have no prior connection to in either the physical or virtual worlds; in other words, people can send unsolicited—and unwanted—e-mails.

The cost to spammers of sending spam is roughly $80 per million messages; by comparison, a direct mail marketing campaign costs roughly $1.39 per recipient, and a telemarketing campaign costs roughly $0.66 per person contacted (Kanich et al., 2008). These discrepancies can be even more extreme when it comes to international bulk communications. While high international postage and calling rates have effectively eliminated the risk of international unsolicited, bulk direct mail or telephone campaigns, sending spam to or from foreign countries incurs no additional costs for the spammer. This low sending cost means that spammers can produce profits even when response rates are extremely low, as they typically are in the case of spam. Direct mailers require an estimated response rate of about 2 percent to make money off a postal mail campaign, but spammers can profit from response rates of less than 0.00001 percent. Spammers can—and do—generate profits from this tiny fraction of responses, but their margins of profit are relatively small. One study on spam conversion rates found that out of 350 million pharmacy campaign e-mails only 28 elicited a sale, suggesting that the profit model for spammers was not especially robust and could be destabilized by even a moderate reduction in spam volume or response rate (Kanich et al., 2008).

Although it is relatively inexpensive to send unsolicited e-mail, there are still substantial costs associated with spam. Economists estimate that the costs of transmitting, storing, and reading unsolicited e-mails total billions of dollars every year, but they are borne primarily by the Internet service providers, e-mail administrators, and recipients, instead of the spammers (Dickinson, 2004). This shifting of costs from the spammers to the spammed and the spam carriers has been the focus of several proposed anti-spam solutions and also a handful of lawsuits seeking to hold spammers responsible for the financial damages incurred by their messages. However, the architecture of e-mail makes it especially difficult to identify spammers and hold them accountable since, in many cases, people are not required to authenticate themselves to send e-mail.

This trait is not unique to e-mail since sending letters and making telephone calls also don't require any identity verification. However, when taken in conjunction with the low marginal cost of sending e-mail, the lack of widespread user authentication schemes contributes significantly to the proliferation of spam, both because it makes identifying spammers and holding them accountable more challenging and because it contributes to the discardability of e-mail address identities (Koomey et al., 2007). Unauthenticated e-mail accounts make it extremely difficult to identify, and therefore to prosecute, spammers who violate legal guidelines, rendering spam regulations, like for instance the United States' CAN-SPAM Act, largely unenforceable. Additionally, unauthenticated e-mail identities increase the challenges associated with constructing reliable mechanisms and blacklists to identify spammers since they can continuously generate new addresses to send messages from, once their old ones are blocked, and can forge addresses from domains and senders that are trusted by the recipient.

As with postal mail and telephone calls, e-mail messages provide a relatively easy means for users to contact people they do not know. However, the effects of this characteristic—like the effects of the lack of user authentication—are greatly magnified by the low marginal costs of sending e-mail. On a given day, someone who might expect to receive 2 or 3 unsolicited phone calls, or pieces of bulk postal mail, could easily receive hundreds of unwanted spam e-mail messages. This ability to send people unsolicited e-mail is particularly problematic given the cost-shifting effect wherein recipients often bear more of the cost of spam than the sender. Some scholars have drawn parallels between spam e-mails and unsolicited faxes, which also produce a cost-shifting effect since recipients bear the cost of the paper the fax prints on. For this very reason, the Telephone Consumer Protection Act (TCPA) of 1991 outlawed the sending of unsolicited faxes, unless the sender and recipient have an "established business relationship." However, while people who receive unsolicited faxes can trace the originating machine relatively easily and therefore hold the offending sender accountable under the TCPA, this option is largely precluded in e-mail by the lack of user authentication.

Taken individually, these characteristics—the low marginal cost of sending spam, the lack of a widespread user authentication scheme, and the ability to easily send unsolicited e-mails to unknown recipients—are not unique to e-mail among all other forms of communication. The combination of all three, however, explains in large part why the e-mail architecture has proved so resistant to anti-spam measures and indicates which traits of e-mail need to be adjusted in order to reduce spam. The remainder of this chapter focuses on techniques that target each of these three traits in ways intended to make sending spam significantly more difficult and dangerous without notably affecting the ability of other users, especially reputable bulk e-mailers, to send e-mail. Exploring those techniques first requires a deeper understanding of the different actors and points of control involved in the e-mail process and the ways in which they can—and cannot—help stop spam.

3.2 E-Mail Points of Control

Unlike many Internet applications, e-mail is not owned or operated by a single centralized company or organization. This means there is an unusually large, decentralized, and diverse set of control points involved in operating e-mail and curbing spam. The decentralization and diversity of these control points has both positive and negative ramifications for spam reduction efforts—positive because there are more points at which spamming can be made difficult in a variety of different ways, but negative because anti-spam measures often require coordination across these numerous different actors. The trajectory of an e-mail message from sender to recipient offers a clearer look at who these different actors are and the roles they place.

E-mail is typically relayed across computer networks using two standard protocols: Simple Mail Transfer Protocol (SMTP) to send messages and Internet Message Access Protocol (IMAP) to receive messages. After a user composes an e-mail message and presses "send," their e-mail client connects to the sender's designated SMTP server on port 25. The SMTP server then looks at the domain name of the recipient address (i.e., the portion of the e-mail address following the @ symbol) and determines whether it lies within its own domain, that is, whether or not the sender and recipient e-mail addresses have the same domain name. If they do have the same domain, then the SMTP server does not need to forward the message to another server and can instead send it on directly to either an IMAP or a Post Office Protocol (POP) server for delivery. If the recipient address domain is different from the sender's, however, the SMTP server queries the Domain Name System (DNS) to retrieve the IP address of the recipient domain name. The DNS returns to the SMTP server a mail exchanger (MX) record, which specifies a preferential list of host names for servers accepting incoming mail for the recipient's domain. The SMTP server then tries to send the message to the most preferred host listed in the MX record, and continues down the list if that transmission fails. After an e-mail reaches the appropriate SMTP server for its recipient domain name, it is usually passed on to either a POP or IMAP server for delivery. For users who rely on POP servers, messages are generally downloaded directly onto their computers and stored there. For users who rely on IMAP delivery, their e-mail messages are stored on their IMAP servers and retrieved by using local clients on their personal computers.

Thus, delivering an individual e-mail message involves several distinct technical components, each of which corresponds to a different person, group, or organization involved in the e-mail sending process. These components include:

1. The e-mail application (e.g. Thunderbird, Exchange, etc.), which is created by software application designers and operated by individual users.

2. The SMTP server (e.g. outgoing.mit.edu or smtp.comcast.net), which is maintained and operated by the host organization or company (e.g. MIT or Comcast).

3. The DNS, which is maintained and operated by the Internet Corporation for Assigned Names and Numbers (ICANN), a non-profit organization incorporated in California.

4. The Internet service provider, such as Comcast or Verizon, which is responsible for transmitting messages between the client and the sending SMTP server, the sending SMTP server and the DNS, and the sending and receiving SMTP servers.

5. The receiving SMTP and IMAP/POP servers (e.g. mail.mit.edu or mail.comcast.net), operated by the individual, group, or organization responsible for the domain name of the receiving address.

6. The recipient users and their e-mail clients (or applications) which retrieve incoming messages from the IMAP/POP servers.

This list offers a basic framework for analyzing and evaluating the different points of control involved in sending e-mail, and subsequently the different actors with a stake in the process. These actors may, in turn, vie for control of different elements of e-mail or, alternatively, face different incentives and forms of regulation.

Though it is useful to divide the actors involved in e-mail into these different categories: users, applications (and application designers), host SMTP, IMAP, and

POP servers, the DNS, and ISPs, it is also important to note that we often see several of these components conflated into a single actor. For example, Comcast is an ISP that also hosts SMTP and IMAP servers. Thus, for users with Comcast e-mail addresses who also rely on Comcast as an Internet provider, their mail server and transmission of their e-mail messages are controlled by a single actor. Alternatively, web-based e-mail clients like Google's Gmail effectively conflate users' e-mail application and SMTP server, both of which are operated by Google. Thus, individual actors can sometimes control multiple of these components, giving them even greater potential to implement more powerful anti-spam tactics, as well as more complicated incentives for manipulating the e-mail process, and greater susceptibility to possible regulation. Despite this growing tendency towards conflation of e-mail actors, e-mail remains, for the most part, a system involving several different components and associated actors, making it a complicated domain to regulate since it is often unclear which actors can or should be regulated and what the long-term implications of those choices might be.

3.3 Existing Legal & Technical Anti-Spam Measures

In December 2003, the Controlling the Assault of Non-Solicited Pornography and Marketing (CAN-SPAM) Act was signed into law in the United States, granting spammers explicit rights to send unsolicited bulk e-mail and prohibiting states from enacting any stricter anti-spam legislation. The law places several restrictions on unsolicited bulk e-mail, forbidding senders from using false headers, harvested e-mail addresses, or open relays (SMTP servers that send mail originating from unknown users) as well as requiring senders to include opt-out links for recipients to unsubscribe from future e-mails, accurate "From" addresses, and relevant subject headings. The primary criticism of the CAN-SPAM Act, and several other similar spam regulations, is that such measures go largely unenforced and have therefore done little to curb the rapid proliferation of unsolicited bulk e-mail (Dickinson, 2004). One survey, done in March 2004, three months after CAN-SPAM was passed, actually found that respondents were "more likely to report that spam has made being online more unpleasant, made them less trusting of e-mail, and reduced their overall use of e-mail" than they had been in June 2003, six months before the law was enacted (Kraut, Sunder, Telang, & Morris, 2005).

The CAN-SPAM Act failed in large part because, in order to hold spammers accountable under it, prosecutors must first identify spammers and then prove that their e-mails violate the specifications of the CAN-SPAM Act, making enforcement cumbersome and time-consuming for the spammed (Soma, Singer, & Hurd, 2008). Section 7 of the CAN-SPAM Act deals specifically with the mechanisms for enforcement of the law and offers some insight into why the legislation may have been so ineffective. Subsection (7)(g) explicitly grants ISPs (or, "Providers of Internet Access Service")—but not e-mail recipients—the right to file civil action suits against spammers who violate the law's stipulations. In March 2004, soon after the bill was passed, a group of four ISPs (AOL, Yahoo, EarthLink, and Microsoft) took advantage of this privilege to file a set of six lawsuits against hundreds of spammers whose messages allegedly violated the conditions of CAN-SPAM. More than six years later, relatively few spammers have been sentenced under CAN-SPAM and the majority of the charges brought by the ISPs in 2004 are still pending.

The CAN-SPAM act is an attempt to regulate spam by regulating the behavior of individual senders (or spammers) but its efficacy appears to be hindered by the fact that enforcement relies on the joint action of individual recipients and their ISPs. An ISP is unlikely to be able to detect whether messages meet the necessary criteria laid out above, for that a recipient will have to read the message and determine whether it includes the requisite opt-out information and content labeling. However, the individuals are not, themselves, able to file class action suits, as designated by CAN-SPAM. Thus, the ISPs and spam recipients are dependent on significant mutual cooperation to enforce the stipulations of CAN-SPAM. Furthermore, some scholars have argued that ISPs do not necessarily have sufficient incentives to be the primary actors empowered to go after spammers. Citing reports that spam adds roughly two dollars per month to the price of individual users' Internet access, Helman (2009) argues:

> One reason Congress created a cause of action for ISPs in the CAN-SPAM Act is the belief that these large industry players would be able to fight spam more effectively than many individual recipients. ...This assumption may not be completely correct. If spam is a problem for all ISPs, they may be able to pass these costs onto consumers without being economically hurt, and therefore the cost of spam may thus be factored into the cost of using e-mail in general. For instance, connection and access fees for data downloaded, either by connection time or by volume of data, are passed directly to the consumer. Although paying to download spam is certainly frustrating to the consumer, it may create an increased profit for the service provider.

Ultimately, CAN-SPAM empowered non-government actors to play fairly minimal roles in regulating spam. The Federal Communications Commission is granted fairly extensive authority to prosecute spammers, but individual recipients have no ability to file lawsuits against spammers and ISPs, though they are empowered to file lawsuits, may be forced to rely on reports from their customers to identify spam and may also face lengthy delays and exorbitant legal fees in pursuing this option.

Another limitation of the Act is that it does not apply to spam sent from other countries, rendering it less effective as spam becomes an increasingly international problem (in 2004, soon after CAN-SPAM was passed, the United States was responsible for roughly 56.7 percent of the world's spam, but by 2008 that percentage had dropped to 14.9). Different national governments have taken slightly different approaches to regulating spam, with the European Union adopting an "opt-in" requirement, whereby senders are forbidden from sending spam unless the recipients have actively opted to receive it. (This requirement is intended to eliminate the "unsolicited" element of bulk e-mail, targeting the third key spam-enabling feature identified above.) A 2004 update to the Dutch Telecommunications Act went even further, outlawing all unsolicited e-mail (Soma et al., 2008). What these strategies have in common with the CAN-SPAM Act is that they have been largely unsuccessful as long-term solutions to curbing spam. In part, this is because spam is an increasingly international issue and national regulations therefore have inherent jurisdictional limitations, but the widespread inability of governments to limit spam suggests that their approach of regulating individual senders may not be the most effective one. Indeed, the combination application-level and ISP-level spam blocking tools appears to have met with far greater success than legal regulation in this area.

At the level of individual recipients and e-mail applications, application designers usually include some functionality to filter out spam into a separate folder from a user's inbox. Users are also often able to mark e-mails as spam using their mail application, signaling to the software that it should try to automatically filter out similar messages. These application-level filters can provide a fairly effective means of reducing how much spam e-mail users are forced to sort through on a daily basis, but they do not alleviate the ISP's burden of transmitting the unwanted messages. Therefore, many ISPs maintain their own programs to filter out spam at the server level and conserve bandwidth for other, more desirable kinds of traffic.

Companies like Google that have condensed the application and server-side elements of e-mail have been some of the most successful at reducing spam. Statistics released by Google in October 2007, summarized in Figure 3-1, showed that users of its popular Gmail reported steadily decreasing levels of spam in their inboxes from April 2004 through October 2007, even as world-wide spam levels increased nearly three-fold. Google automatically filters many messages into a pre-marked spam folder for each user, but also learns from users' individual spam marking habits which messages are widely regarded as unwanted and uses that information to apply to its larger customer base. Thus, while a standard e-mail application might only learn which messages to mark as spam from users' individual spam marking activity, Google is able to pool the spam marking of all of its clients and apply that knowledge to all of its users.

Ultimately, though the amount of spam being sent has not decreased in recent years, there is some evidence that progress has been made in managing and filtering it more effectively. The percent of Internet users who said spam was a "big problem" for them dropped from 25 to 18 percent between 2003 and 2007, according to a report by the Pew Internet & American Life Project. Two possible factors may contribute to this drop: users may be growing more accustomed to, and therefore less irritated by, spam messages, and spam filtering programs may also be improving and becoming increasingly effective. The trend towards conflation of different e-mail actors may contribute to these improved spam filtering results, due to the removal of the distinction between separate e-mail applications and servers. This separation can prevent pooling the information about spam messages identified by multiple different users and lead to potentially less powerful, though more personalized, spam filtering.

Currently, government regulation plays a fairly minimal role in limiting spam and one of the challenges to governments, when considering how to regulate e-mail, is

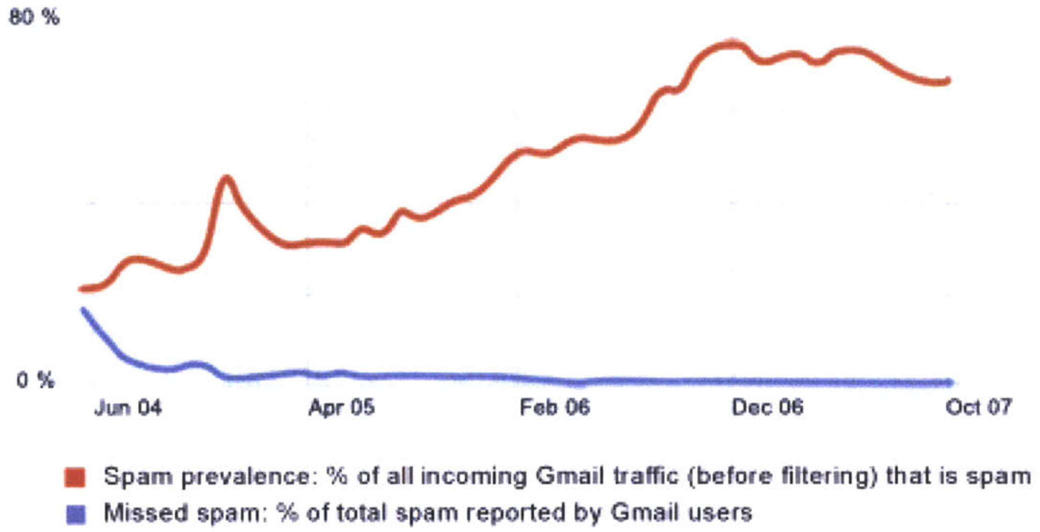


Figure 3-1: A chart released by Google showing the rates of total incoming spam and spam reported by Gmail users.

how many different, independent actors are involved in the process. The convergence of these different actors could provide new opportunities for—as well as new unintended consequences of—spam regulation and government interventions intended to raise the marginal cost of spam to spammers, improve user authentication for e-mail, and increase the risks associated with sending unsolicited e-mail to strangers. The following sections explore different legal and technical approaches aimed at targeting each of these three goals.

3.4 Raising the Costs of Sending Spam

Unsolicited bulk communications existed long before e-mail, in the form of postal mailings and phone marketing campaigns. Though irritating to consumers, these other modes of mass messaging have never approached anything near the volume of spam e-mail, due in large part to the higher costs of using them, shown in Table 3.1. E-mailing 1 million people costs less than $2,000 (primarily to pay for the necessary computing power, usually by renting a botnet) but sending a conventional postal mailing to that many people would cost nearly $200,000 in postage, plus paper and printing costs (Kraut et al., 2005). In light of this discrepancy, several researchers have hypothesized that one effective way to reduce spam mailings would be to institute a form of "e-mail postage" that would raise the marginal cost of spam high enough to render it unprofitable.

One means of instituting e-mail postage is charging users for each e-mail they send, just as people pay to send physical letters. An e-mail pricing scheme could even be designed to mirror that of the postal service, charging users more to send larger messages. However, associating a financial cost with sending e-mail would also

Table 3.1: Cost Per Recipient of Different Marketing Media (Judge et al., 2005)

Medium	**Cost Per Recipient**
Direct Mail	$1.39
Telemarketing	$0.66
Print—targeted	$0.075
Print—general	$0.067
Fax	$0.05
Online Ads	$0.035
Spam	$0.0005

have several drawbacks. Most notably, it would inconvenience all legitimate e-mail users (i.e., those who are not spammers), especially poorer users and legitimate bulk e-mailers, who send solicited messages to large circulation lists. Furthermore, the details of the payment scheme raise several open questions about who the postage fees would be paid to and how would they be processed and monitored. In the past, some ISPs have attempted to profit from similar pricing schemes, directed specifically at bulk e-mailers. In February 2006, AOL and Yahoo both announced they would whitelist (i.e., not filter) all mail from companies willing to pay from $\frac{1}{4}$ of a cent to a penny per message. While the companies both said they would continue to accept and deliver free e-mails, these messages would be subject to the regular spam filters while the paid, or certified, messages would be delivered directly to users' inboxes, without being stripped of images or links, and would be marked as "Certified E-Mail" (Hansell, 2006). The profits from this certification scheme were to be split between the e-mail providers and Goodmail Systems, a company that administered a certification process for high-volume senders before it shut down in February 2011, after failing to find a market for certified e-mail and being criticized by numerous non-profit organizations which maintained large e-mail distribution lists (Atkins, 2011). The non-profit political advocacy organization MoveOn denounced Goodmail's certification model with particular vehemence, posting on its website that "Charities, small businesses, civic organizing groups, and even families with mailing lists will inevitably be left with inferior Internet service unless they are willing to pay the 'e-mail tax' to AOL."

Given the difficulty of instituting a reliable micropayment scheme, the fear that such a scheme might discourage legitimate e-mailing, and the risk of creating another point of failure for e-mail—if the micropayment system fails, the e-mail architecture crashes—anti-spam researchers have also investigated other forms of payment besides money. Just as users can create costly virtual identities by investing either money or time in them, senders could be required to invest "time" in sending their e-mails, rather than money. These "pricing via processing" schemes force a machine to perform a simple computation requiring a certain number of CPU cycles each time an e-mail is sent from it (Dwork & Naor, 1993). In the case of an average e-mail user, this cost might be trivial, but for spammers sending out millions of messages at a time it could effectively cripple their operations by dramatically slowing down the sending process. Proposed variations on the processing pricing schemes include using a Turing test to ensure that messages are being sent by a person and memory-bound functions that might be more equitable than the CPU cycle-based-computations, since memory access speed varies less than CPU speeds across different machines. These types of processing price schemes could be both more feasible to implement reliably and less likely to inconvenience regular users than standard monetary prices, however they would still pose a major obstacle for legitimate bulk e-mailers. Dwork and Naor (1993) have proposed addressing this problem by embedding a short cut or "trap door" in the e-mail pricing function so that "given some additional information the computation would be considerably less expensive" for the sending of legitimate bulk mail, such as a call for papers for an academic conference. This solution also exhibits some of the drawbacks of the certified e-mail system, however, by introducing yet another point of control—in this case, the system manager who would determine which bulk mailings were legitimate—which might be susceptible to failure or corruption.

Pricing schemes aim to make the cost structure of e-mail more closely resemble those of postal mail and telephone calls, so that senders bear more of the expense of bulk mailings and can less easily shift those costs onto recipients and intermediary parties. Paying to send e-mail also has a potential signaling value to recipients, by informing them about the value of the message to its sender. Just as a user can invest more or less in an online identity as a signal to other users about how reputable or valued it is, an e-mail sender might be able to pay a higher delivery fee or bond price to signal to the recipient the importance of the message (Kraut et al., 2005). Signaling in this manner and imposing higher costs on senders can play a role in reducing spam, but these methods also present some serious drawbacks. In many cases they may disadvantage all bulk e-mailers, not just spammers, restricting even non-commercial, solicited mailings. Additionally, it is not clear that imposing some sort of monetary or processing cost will necessarily be sufficient to defeat spam—even spam campaigns whose costs are primarily born by the spammers, such as those by telephone, have become sufficiently irritating and numerous to warrant the institution of federal do-not-call lists (Dickinson, 2004). Therefore, it is worth examining the other characteristics contributing to spam's profitability besides its low marginal cost to senders: the lack of user authentication and the ease with which users can send messages to strangers.

3.5 Improving User Authentication

In 1982, when SMTP was first written for the relatively small number of academic, online users, its authors had no reason to be worried about the proliferation of e-mail spam and therefore no reason to design their protocol to authenticate users or guarantee the integrity of the messages they sent. Accordingly, the only means embedded in the protocol for tracing the source of an e-mail is the address in the "From:" field, which can be easily forged by the sender—some estimates suggest that as much as two-thirds of all e-mail sent uses spoofed sender addresses—and the "Received:" headers added by each host that relays the message (Goodman, Heckerman, & Rounthwaite, 2005). These headers list the name and address of both the system relaying the message and the system it received the message from (Dickinson, 2004). Spammers cannot stop these headers from being added by the intermediary hosts, but it can still be difficult to use the headers to identify the originating source of an e-mail, both because such attribution requires very thorough examination of the headers and because spammers can use open relays and fake headers to further complicate the tracing process.

By sending spam through open relay servers, spammers can avoid identifying the sending computer, so it is often impossible to trace a message back from that relay to its originating machine. For this reason, open relay sites are often blacklisted by ISPs and using them to send e-mail is outlawed in the United States by CAN-SPAM. Additionally, spammers can add extra, fake headers to messages in order to confuse efforts to trace the originating source. These false headers can, potentially, be identified by working through each header and verifying its authenticity with an administrator at each intermediary, but this process is often prohibitively time-consuming (Dickinson, 2004). In short, there is no easy and reliable way to trace the origin of spam e-mail messages, and even in instances where the source mail server is successfully identified it may still be difficult to locate the specific user responsible for sending a certain message. This inability to effectively identify spammers makes it significantly more difficult to hold them accountable for their actions under a legal framework. Dickinson (2004) writes, "Spam can often be filtered or blocked, but the underlying architecture of e-mail provides an effective barrier between law enforcement and the perpetrators of spam."

One alternative, or potentially complementary, approach to imposing fees on e-mail that more specifically targets this barrier would be instituting a system of user authentication certificates for e-mail senders. Under such a system, senders would digitally sign their messages and recipients could then indicate which sender certificates they trusted to receive mail from and which ones they did not. Once again, this solution creates new points of control with a considerable influence over the e-mail application: in this case, the group or organizations responsible for administering certificates. Some scholars (Dickinson, 2004) have advocated for a government-administered certificate authority, arguing that this would be the most reliable means of associating a valid name and identity with an online certificate. Concerns that this would too drastically diminish the privacy and anonymity afforded by the Internet have given rise to proposal for more informal systems of self-signed certificates that might less reliably identify a specific person but could serve a similar purpose in allowing recipients to accept or reject messages from other users. Still other schemes would allow for domain authentication, using digital signatures to ensure that e-mail is sent from a specific domain, while still maintaining some of the sender's anonymity.

Inevitably, widespread user authentication would impact the ability of e-mail senders to protect their anonymity though, depending on the implementation, it might still be possible for senders to conceal their real identities from some parties. For instance, such schemes could be designed either to depend on third-party enforcement or to convey identity between the message end-points; in the latter case, senders could potentially reveal their identity more narrowly, only to the receiver instead of to additional third parties. Regardless of implementation, however, the entire point of such an anti-spam mechanism would be lessening senders' ability to conceal their identities from enforcement efforts, thereby reducing anonymity. However, if combined with some of the pricing mechanisms described in the previous section, an authentication mechanism for e-mail could play an important role in helping users decide for themselves what elements of e-mail they most highly valued. For instance, authenticated users who digitally signed their e-mails might be permitted to send their messages for free, without paying any financial or processing price, since service providers and recipients could feel more confident in their ability to hold an authenticated sender accountable for any unwanted spam. Alternatively, senders who valued the anonymity of their e-mail addresses more highly than the ability to send free messages could forego the authentication process but instead pay a fee of some sort to demonstrate the value of their messages. In this manner, users could be permitted to make their own individual trade-offs between how much anonymity and accountability they wanted to embed in their e-mail identity, adjusting that balance for different addresses or even different messages sent from the same account. A dual pricing-authentication system would allow legitimate bulk mailers to send their mailings for free by submitting to an authentication process and also enable parties with compelling reasons to maintain the anonymity of their e-mail identities to pay slightly more for that privilege. Similarly, individual users could customize the settings for which types of e-mail they wished to receive: only messages from authenticated users they know and trust, or messages from all authenticated users, or messages for which a postage fee has been paid, or even un-authenticated, un-paid messages, as well. Different users, after all, may have different definitions of spam and different levels of tolerance for it, so it is often easier to let them decide which types of e-mail they want to receive than to impose a universal, or even national, definition of what does and does not constitute spam.

3.6 Increasing the Risks of E-Mailing Strangers

While authentication schemes would enable users to indicate which certificates they do and do not trust, there could still be a substantial amount of spam e-mail sent from unknown addresses and domains. Though users could elect not to receive, or read, messages from unknown certificates, they might decide against this course of action for fear of missing some legitimate mail. There could be times where it would be valuable to users to receive messages from some strangers, without having to read every unsolicited spam message directed to them. In these case, it might be beneficial for users to be able to distinguish how much strangers had invested in sending them unsolicited messages. Microsoft's Penny Black project focuses on precisely this problem and sums up the main idea behind its research as: "If I don't know you, and you want to send me mail, then you must prove to me that you have expended a certain amount of effort, just for me and just for this message."

To increase the financial risks of sending strangers unsolicited e-mails, some proposals have called for the creation of a payment scheme involving "sender bonds," in which senders commit to spend a certain amount of money in a bond that accompanies each e-mail message they send. "After viewing the message, the recipient can then choose to claim the bond value or not—seizing the bond makes the spammer pay for demanding attention, even if it's only the few seconds needed to delete a message," Koomey et al. (2007) wrote in a *Wall Street Journal* op-ed advocating this approach. "If even a small percentage of recipients claim the bond, spam campaigns will become uneconomic," they explained.

This system is related to the micropayment "e-mail postage" schemes but makes more of an attempt to distinguish between solicited and unsolicited mail and specifically discourage the latter. Bonds also have the benefit over e-mail postage fees of reducing the expense and inconvenience to e-mail users who are not spammers since they could avoid paying most—if not all—fees by only sending messages to recipients who were likely not to claim the bond. The e-mail bond system also has several of the same weaknesses as the pricing schemes, however, in that it, too, would require a secure system of micropayments and could potentially encourage corruption and theft. Such a bonding set-up could still potentially incur some cost to everyone who sends e-mail, thereby discouraging use of the medium. Furthermore, it is entirely reliant upon the implementation of secure, easy-to-use online micropayments, which some researchers have argued are unlikely ever to play more than a marginal role in the Internet economy (Odlyzko, 2003). However, the principle of the bonding scheme—that users are not subject to any accountability cost or punishment unless they engage in a discouraged activity like spamming—is a useful one when considering online accountability mechanisms and can also be applied in different ways to many applications beyond e-mail.

3.7 The Future of Fighting Spam

Since, on their own, both technical and legislative approaches have proved unequal to the task of curbing spam, a truly effective anti-spam framework would likely require much greater and more thorough coordination between these legal and technical measures, to ensure that each can mitigate the weaknesses of the other and, in turn, bolster each other's effectiveness (Sorkin, 2001; Dickinson, 2004; Lessig, 2006). Accordingly, the most promising anti-spam methods seem to be those that target spam on several levels and leave as much choice and control as possible to the end-users and ISPs. For instance, widespread implementation of an authentication mechanism could be coupled with a processing pricing scheme, so that e-mail senders could choose either to digitally sign their e-mails or to perform a calculation before sending—or to do neither, or both. Users and ISPs could then decide whether they wanted to transmit or receive messages that had neither a digital signature nor a computational price associated with them. Such a scheme would still require instituting a large-scale, secure micropayment infrastructure, thereby creating another potential point of failure for e-mail, but it would also provide senders and recipients some choice in what kind of e-mails they wished to send and receive. For instance, legitimate bulk e-mailers could send digitally signed messages and avoid the cost of intensive computation, while users wishing to remain anonymous could instead pay the "processing" fee for their messages. Notably, neither of these systems need be mandated by law since, if they achieved widespread, international popularity, it is likely that messages without either a digital signature or a processing fee would be increasingly dropped by ISPs or filtered out by end-users. This hybrid system has the additional benefit of being easily extended to e-mail from all countries, thereby countering the jurisdictional limitations of a strictly legal approach.

Levchenko et al. (2011) have proposed an alternative means of crippling the profit model for marketing spam: cutting off spammers' ability to receive credit card payments by going after the banks that process these payments for them. They found that the same three banks were used for processing payments for more than 95 percent of the items they found advertised in spam e-mail campaigns. Since so few banks are willing to process these high-risk card-not-present transactions, if prominent Western credit card companies like Visa and Mastercard refused to settle such transactions with the identified spam-supporting banks it could potentially destabilize the spam economy. The time and money needed to establish a business relationship with a new bank is sufficiently substantial that the researchers hypothesized a "financial blacklist" of banks that support spam transactions could be updated more quickly than the spam merchants would be able to find new banks, creating a "rare asymmetry favoring the anti-spam community."

Exploiting the banking control point as a means of reducing spam is another approach to trying to adjust the system of e-mail to favor legitimate users over spammers, though it provides no protection against non-commercial spam or phishing attacks. Like the other approaches discussed in this chapter, this method has strong advantages in targeting a very specific—and very common—type of spam without harming legitimate users but it also has clear disadvantages in only targeting commercial spam and requiring buy-in from powerful credit card companies and financial institutions who might fear losing business by cooperating with these measures. Similarly, every anti-spam strategy discussed here has flaws and drawbacks. Authenticating users may enable better anti-spam enforcement but it will also reduce the anonymity of e-mail identities; charging e-mail postage fees could reduce spam but could also reduce legitimate, non-spam e-mail use; instituting bonds may make it harder for spammers to send unwanted, unsolicited e-mails but may also discourage other users from initiating contact with strangers via e-mail, driving them instead towards other modes of communication. These trade-offs emphasize the importance of combining and coordinating anti-spam approaches so that users may establish for themselves, based on their own e-mail identities, where they wish to fall on these spectrums: how anonymous or identified they want their e-mail addresses to be, how much they do or don't want to pay to send e-mail, how wary or eager they are to receive messages from people they do not know.

Lessig (2006, p. 264) writes:

> One great virtue of e-mail was that it would lower the costs of social and political communication. That in turn would widen the opportunity for

> political speech. But spam-blocking technologies have now emerged as a tax on these important forms of social speech. They have effectively removed a significant promise the Internet originally offered.

Restoring that promise and ensuring the sustained utility and value of e-mail communication requires stronger strategies that more directly cut at the spam profit model, more effectively coordinate legal, social and technical approaches, and more fundamentally reconfigure the architecture of e-mail to favor the spammed over the spammers.

Chapter 4

Second Life

> The bride and groom prepare to walk down the aisle. They have been lovers for over a year in Second Life but have never met in the "real" world. In fact, they have not shared any information about their real-world lives—the bride might be a man, the groom a woman, either might already be married in the "real" world—but you feel genuinely happy as they exchange vows.
>
> —Tom Boellstorff, *Coming of Age in Second Life*

Weddings are not uncommon in the virtual world of Second Life; the community's extensive online economy even features a variety of bridal gowns, wedding planners, and even elaborate multi-tiered cakes available for purchase. Despite all these trappings of real-world marriages, Boellstorff (2010) notes that married couples in Second Life know each other only through pseudonymous avatars, they have made a conscious and purposeful decision to conceal from each other the details of their actual identities, usually choosing instead to create entirely separate and unlinked identities—and relationships—within the online universe.

This emphasis on anonymity and pseudonymous identities is not unique to Second Life among online communities. Many other applications that similarly aim to create communal settings in which large groups of users can interact with each other also rely on a culture of pseudonymity. These communal platforms include applications like Wikipedia, the popular user-written and edited online encyclopedia, LambdaMOO, one of the early text-based online communities, and the popular fantasy game World of Warcraft. Certainly, there are some users of all of these platforms, including Second Life, who are perfectly happy to reveal their real-world names and identities within the online communities but, in general, they are the exception rather than the rule. What sets these communities apart from social networks like Facebook and Myspace, and even, to a lesser extent, review sites such as Yelp, is that the large majority of Second Life users do not want their online identities within these communities to be in any way associated with their actual identities.

It is this type of community, where pseudonymous identities are the norm and users expect and intend that other members will be unable to link their avatars back to their real-world identities, that will be the focus of this chapter. Specifically, this chapter will look at the identity schemes and regulatory mechanisms implemented in one of these communities, Second Life, the immensely popular online world launched in June 2003 by the California-based Internet company Linden Lab. With a bustling online economy based on "Linden dollars" (which can be purchased with and exchanged for real money) and numerous opportunities for users to engage with the virtual world, by building houses, launching businesses, socializing with other avatars, and exploring the vast, graphical "grid" landscape, Second Life boasts roughly 1 million active users and a GDP of $567 million, according to a 2009 estimate.

A community involving so much activity—and so much money—has inevitably seen a certain amount of misbehavior over the course of its development, and indeed, in the eight years since its launch Second Life has been host to everything from organized "griefing" groups like the infamous Patriotic Nigras, to identity theft, copyright infringement cases, denial-of-service attacks, and virtual riots. Many, though not all, Second Life users provide Linden Lab with their credit card numbers when creating their accounts, however this link to their real identities is rarely, if ever, used to hold Second Life users accountable for their avatars' misbehavior. This is partly because tracing avatars to their real-world counterparts requires the involvement of Linden Lab, which actively tries to distance itself from moderating and regulating the day-to-day social conflicts within Second Life, and also because many of Second Life's users consider anonymity to be an essential element of the community.

4.1 Anonymity & Identity Schemes in Second Life

Why is anonymity so important to Second Life users? Arguments about the value of anonymous Internet activity to protect political dissidents and protestors governed by authoritarian regimes have relatively little relevance to Second Life, which is not a popular platform for meaningful social protest. But while anonymity may not be used in Second Life specifically to facilitate free speech or political uprising, this does not make its users any less committed in the idea that their Second Life avatars should be firmly divorced from their real-life identities. In September 2006, Linden Lab released a security bulletin announcing that one of their databases had been breached, potentially giving the intruder access to the unencrypted names of Second Life account holders and thereby allowing them to link those names to the corresponding users' avatars. Though Linden Lab assured its customers that their credit card numbers were completely protected from the hackers, many users were still outraged by the possibility of their anonymity being compromised, even if their bank accounts were safe. In another example of how much Second Life residents value their anonymity, there were widespread negative user reactions to the addition of voice chat functionality, which was introduced in Second Life in 2007 despite protests by many users who feared the vocal interactions would "damage a border between the virtual and actual that they wished to maintain" (Boellstorff, 2010, p. 113).

Some Second Life users may have more obvious motivations for wanting to maintain this border; for instance, those whose avatars have active careers as prostitutes might fear that such activities would be embarrassing and even personally or professionally damaging if linked back to their actual identities. Others may want the opportunity to experiment with personality traits or characteristics—ranging from age to gender to sexual preferences—that are radically different from those of their actual identities. Still others may simply desire to keep their two worlds separate for no reason other than to be able to create and foster multiple identities. Understanding these motivations for user anonymity provides some insight into the costs associated with the loss of this anonymity which, in turn, helps situate Second Life along the axis of differing degrees of online anonymity. If the primary costs, or consequences, of losing one's anonymity in Second Life are embarrassment and reputational damage then the corresponding identity scheme may not require quite so complete a degree of anonymity as schemes implemented in platforms used by political dissidents and protestors, for whom the cost of having their true identities revealed could be as high as imprisonment or execution.

This concept of different tiers or degrees of anonymity is central to understanding the full range of possibilities afforded by the online identity space. Donath (1999, p.53) explains:

> In the virtual world, many degrees of identification are possible. Full anonymity is one extreme of a continuum that runs from the totally anonymous to the thoroughly named. A pseudonym, though it may be untraceable to a real-world person, may have a well-established reputation in the virtual domain; a pseudonymous message may thus come with a wealth of contextual information about the sender. A purely anonymous message, on the other hand, stands alone.

Second Life, like most online communities, falls somewhere in the middle of the spectrum Donath describes: it is not "purely anonymous," nor is it "thoroughly named." Instead, it provides users with a pseudonymous identity scheme that in many, though not all, cases can be linked back to users' real-life identities via their credit card information, but only by Linden Lab. Both of these elements—the pseudonymous avatars and the possibility of identifying their actual-world counterparts—play an important role in defining Second Life's identity scheme and determining where it falls within the proposed accountability-anonymity axes.

In Second Life, users create avatars which they can name and assign physical attributes to as they choose. Notably, users are free at any time to change any trait of their avatars, except for the screen name and date of the avatar's creation (Boellstorff, 2010). Thus, even though a user may decide to completely transform his avatar every day—switching between identities as diverse as, say, a five-year-old boy, a sixty-five-year-old woman, and a twelve-year-old greyhound dog—the avatar's name provides some degree of continuity in the avatar's identity and reputation within Second Life. This continuity is an essential element for the creation of communities (online and otherwise) (Donath, 1999) argues, writing, "The motivation for many of the qualities we associate with community, from cooperative behavior to creative endeavor, depends on the existence of distinct and persistent personae." Of course, it is possible for individual users to own and operate multiple avatars with different screen names, but even then, each individual avatar is a "persistent persona" and can be associated with its previous actions and traits in Second Life by other users. Furthermore, surveys of online communities show that even those users who do have multiple characters typically report using one primary identity most of the time (Schiano & White, 1998).

Thus, the two unchangeable traits of each avatar—its name, and, perhaps to a lesser degree, its date of creation—serve as the basis for a pseudonymous (but not anonymous) reputation system within Second Life. "Screen names are by definition not anonymous, and the virtual selfhoods tied to them have become increasingly consequential," (Boellstorff, 2010, p. 122) points out. In the early versions of Second Life, Linden Lab implemented some tools to help users codify reputation and rating schemes associated with other avatars. Initially, for instance, any resident could pay L$1 (one Linden dollar, or approximately $.003) to rate another avatar positively or negatively. The cost of these ratings was so low, however, that users began to organize "ratings parties" the entire purpose of which was to shower a single user with vast numbers of either positive or negative ratings to skew their virtual reputation. In October 2005, Linden Lab responded to these parties by raising the cost of a positive rating to L$25 and removing negative ratings altogether. After this price increase, however, users ceased to make any regular use of the ratings system and it was discontinued in April 2007 (Boellstorff, 2010). Though Second Life no longer has a formal ratings-based reputation system in place, the informal, or implicit, means of building reputation by observing and discussing other avatars' behavior remains integral to the virtual world. Meanwhile, other online applications, like Yelp and Wikipedia have met with more success in developing formal reputation systems to enable their users to assess the pseudonymous identities of other members of the community. For instance, Yelp user profiles—discussed in greater detail in chapter 6—display not just the name and creation date associated with an account, but also the number of reviews that have been authored under that account, the distribution of ratings given by that user, the number of other users who have rated that account holder's reviews as helpful or unhelpful, and a variety of other statistics that create a fairly rich reputation, or context, for other Yelpers to take under consideration when reading reviews.

Second Life avatar profiles are largely determined and created by individual users, however, there are certain fields automatically populated by Linden Lab that create some reputational elements. The date of account creation may serve this purpose to some degree, but a potentially richer signal to other users may be the information provided about whether or not the account holder has provided Linden Lab with "identity verification" in the form of a valid credit card account. Initially, when it was launched in 2003, Second Life offered users a choice between a free account and a paid "Premium Account." Premium account holders were permitted to own land while free account holders were not, but even those users who elected to have a free account were required to enter a valid credit card account upon registering as a form of "identity verification." This meant Linden Lab had the capability to trace each avatar back to a real-world credit card holder, placing some limits on the degree of anonymity (or pseudonymity) of the community, as well as the corresponding disinhibition residents might experience in Second Life.

In June 2006, Linden Lab decided to change its registration requirements so that users wishing to create free accounts would no longer have to provide a credit card number, or any other identifying information beyond a username and password. This change was intended to broaden Second Life's user base, since Linden Lab had noted that roughly half of the people who initiated the registration process stopped at the stage where their credit card information was requested. Indeed, the new procedure led to an approximately fourfold increase in the rate of new users signing up for accounts, but many older residents protested the change, joining together to create Proposition 1503 which lobbied Linden Lab to reinstate an identity verification requirement for new registrants (Boellstorff, 2010).

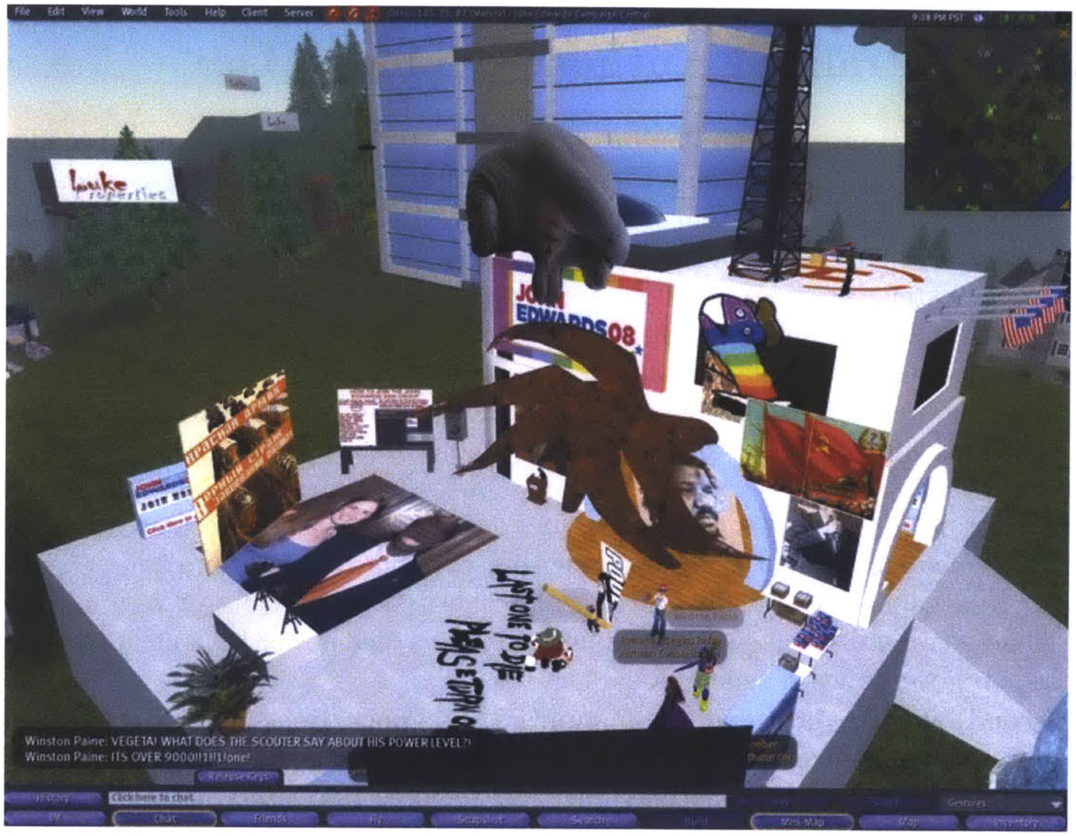
The Second Life users behind Proposition 1503 were concerned that the greater degree of anonymity afforded by these unverified free accounts would be detrimental to the community's accountability mechanisms and might enable the new users to engage in more irritating or damaging forms of behavior without fear of any repercussions. In response to these concerns, Linden Lab issued an announcement later that same month: "Each resident's profile now includes a field revealing . . . one of three status entries: (1) 'No Payment Info on File'—account was created with no credit card or Paypal; (2) 'Payment Info on File'—account has provided a credit card or Paypal; (3) 'Payment Info Used'—credit card or Paypal on account has successfully been billed. We plan to provide features in future updates to mark specific parts of the Second Life world (or allow residents to mark their own land) as accessible only to accounts with payment information" (Boellstorff, 2010, p. 235). In this manner, account holders were still provided with the same option to provide credit card information, but this choice now became a public piece of their online identities and could be used by other residents to assess their avatars, as well as by Linden Lab to limit their privileges within the virtual world.

Thus, Second Life's identity scheme has been revised by Linden Lab over time to better enable users to choose what degree of anonymity and privileges they want for themselves. Residents who value anonymity very highly can elect not to provide Linden Lab with any identifying credit card information, and while these avatars can still be held accountable for their actions using some of the social norm and community policing mechanisms described later in this chapter, their real-world operators are clearly less accountable to Linden Lab and, consequently, they have fewer privileges to do things like own land or access certain areas. On the flip side, those users who value these privileges more highly than they do their complete anonymity, can enter credit card information and obtain the wider array of capabilities that are associated with this higher degree of accountability (and, in some cases, higher financial cost) while still enjoying the relative anonymity of a community in which none of the other users can trace their avatars back to their actual identities. Every Second Life avatar can therefore be more or less anonymous, more or less accountable, more or less privileged, and more or less expensive, depending on the preferences of each individual user. This is not to say that Second Life's pseudonymous identity scheme is without its share of problems, or that individual end-users possess all (or even most) of the power when it comes to regulating and mitigating these problems. Inevitably, with a user base of one million people who can avail themselves of differing degrees of anonymity, Second Life has been host to a number of avatar conflicts and hacking attacks over the course of its eight-year history that Linden Lab has struggled to keep under control by leveraging a variety of different control points.

4.2 Misbehavior & Points of Control in Second Life

LambdaMOO founder Curtis (1997, p. 129) observed that "the most significant social factor in [online communities] is the perfect anonymity provided to the players." Though this anonymity is far from "perfect" in most cases, it is certainly true that the perceived boundary between users' real-world and online identities can radically transform and even eliminate some of the standard social inhibitions we take for granted in the actual world. As Reid (1999, p. 112) puts it: "Users experience a redefinition of social inhibitions; they do not experience the annihilation of them." Second Life proponents and residents are quick to point out that these redefinitions may have positive ramifications as well as negative ones, whether by allowing users to express elements of their personality they would otherwise be too shy or scared to display in the actual world, or even by encouraging an unusual degree of altruism and generosity towards other residents (Boellstorff, 2010). But, as Curtis (1997, p. 130) notes, "This protective anonymity also encourages some players to behave irresponsibly, rudely, or even obnoxiously ... In general, such cruelty seems to be supported by two causes: The offenders believe (usually correctly) that they cannot be held accountable for their actions in the real world, and the very same anonymity makes it easier for them to treat other players impersonally, as other than real people."

In online communities like Second Life, these malicious behaviors fall broadly into two main categories: "in-world" griefing attacks, intended specifically to disrupt other users' enjoyment of the virtual community, and hacking attacks on the platform's technical infrastructure, which also aim to ruin other users' experience but do so by targeting the community's code framework, rather than operating within the social confines of the community. In Second Life, examples of the former have included everything from taunting users with insulting epithets, to simulated sexual harassment, building large floating boxes right next to neighbors' windows and then demanding that they pay exorbitant fees to have the boxes removed and restore their views, copyright infringement (usually in the form of selling copies of products created by—and therefore the design of which belongs to—other avatars), and a much-publicized 2007 attack on the Second Life campaign headquarters of presidential candidate John Edwards, shown in Figure 4-1. During this attack, according to a post on Edwards' campaign blog, "a group of Republican Second Life users, some sporting Bush 08' tags, ... plastered the area with Marxist/Leninist posters and slogans, a feces-spewing obscenity, and a Photoshopped picture of John in blackface, all the while harassing visitors with right-wing nonsense and obscenity-laden abuse" (Wheaton, 2007). The hacking or code-based type of misbehavior usually takes the form of denial-of-service attacks, often perpetrated by avatars who create self-replicating objects within Second Life that continue copying themselves until they ultimately crash the Linden Lab servers.



Winston Paine: VEGETA! WHAT DOES THE SCOUTER SAY ABOUT HIS POWER LEVEL?!
Winston Paine: ITS OVER 9000!!!1!!1one!

Figure 4-1: Screenshot of 2007 attack on John Edwards' Second Life campaign headquarters (Brownlee, 2007).

All of these forms of misbehavior adhere loosely to the notion of griefing: in each case, the griefer's action is intentional and the griefer derives some joy (or satisfaction) from it, while, at the same time, it results in other users deriving less enjoyment from the online community (Boellstorff, 2010). However, the distinction between those attacks which target the technical infrastructure owned and operated by Linden Lab and those which do not is an important one because it begins to get at the question of what different control points exist in Second Life and which ones can act to deter or mitigate different kinds of detrimental behavior. In the case of hacking attacks, like the ones triggered by self-replicating objects, Linden Lab is the only point of control capable of containing and stopping the attack, as well as holding the attackers accountable. In-world harassment and assault, on the other hand, can be addressed, mitigated, and in some cases even punished by a variety of other actors, including individual avatars and end-users as well as self-organized forms of communal policing and governance. The following sections will deal with each of these three points of control—Linden Lab, individual end-users, and mid-level community governance organizations—looking at the capabilities and limitations of each one, with a focus on what types of attacks each control point is best equipped to deal with as well as what accountability mechanisms each one can effectively implement.

4.3 Centralized Authority: Linden Lab

Doctorow (2007) argues that "online games like World of Warcraft and Second Life are absolute dictatorships, where the whim of the companies controlling them is law." On the one hand, it is true that in many ways Linden Lab is able to do pretty much as it pleases in governing and adjusting Second Life (though it is doubtless guided to a great extent in these decisions by concerns about satisfying and retaining its users), on the other hand, though, there are a variety of other, complementary governance mechanisms possible in online communities like Second Life. These additional mechanisms are often useful because there are numerous types of misbehavior that a centralized moderator or authority, like Linden Lab, may actively choose not to involve itself in, either because it does not have the necessary time and resources, or because it feels it is in some sense not their place. This section examines the governance capabilities of Linden Lab within the context of Second Life—and more broadly, the capabilities of any centralized authority that controls, owns and operates an online community—as well as the limitations on this power that necessitate other additional governance schemes.

Second Life, like most online communities and applications in general, is governed by a "Terms of Service" (ToS) agreement that all users must sign upon joining. Authored by Linden Lab, the Second Life ToS requires that users uphold community standards and refrain from any of the "big six" transgressions of intolerance, harassment, assault, disclosure (of personal information), indecency, and disturbing the peace (Boellstorff, 2010). However, although these norms were codified in the company's ToS agreement, the six transgressions listed were not necessarily well-suited to being governed or punished by Linden Lab. Users who witnessed ToS violations had the option of filing "Abuse Reports" with Linden, but Boellstorff (2010) found that only about 6.5 percent of residents regularly did so and many of those who did experienced long delays and minimal, if any, response.

When Linden Lab did choose to exercise its power to punish a Second Life resident on the basis of reported abuses, the company's responses could range from issuing warnings to the accused users, to temporarily or permanently suspending their accounts (Boellstorff, 2010). Though users could relatively easily create new, free accounts following such suspension, these accounts (by virtue of being free) would be more limited in their privileges and would also lack any reputation, money, or property built up by the misbehaving users' previous avatars. This in itself might be sufficient to deter some users from risking account deletion, while Linden could implement additional restrictions—such as blocking the credit card accounts associated with suspended accounts from being used to create new ones—to further guard against particularly problematic users abusing their ability to create new identities.

For more mild offenses, which were not punished by account termination, Linden Lab operated a "police blotter" on the Second Life website, listing their recent enforcement activities as a means of shaming offending users and signaling their behavior to others in the community (Boellstorff, 2010). At the opposite end of the spectrum, in response to the most extreme transgressions, such as denial-of-service attacks on the Second Life grid or attempts to hack the Linden servers to obtain credit card numbers and other actual-world information about users, Linden Lab would sometimes refer the case to government law enforcement authorities. In these instances, Linden might choose to use its database of credit card account records to link an avatar back to an actual-world person and hold that person accountable to actual-world laws and standards, rather than simply trying to hold their avatar accountable to cyberspace norms and standards. These cases were the exception rather than the rule, however, and in many of them, especially when the offending party lived outside the United States, law enforcement remained extremely difficult, even if the guilty individual had been identified via Linden Lab's credit card records. Still, Linden Lab's ability to trace avatars back to actual identities is notable simply because it is the only control point within Second Life that has this power. Donath (1999, p. 54) notes that, "The sanctions to offensive online behavior can be roughly divided into two main categories: those that involve making a connection to a real-world person and those that do not." While accountability mechanisms for pseudonymous personae can be—and are—implemented at all levels of governance within Second Life and similar online communities, sanctions requiring a connection to avatar's real-world counterparts can only be implemented by Linden Lab. For the most part, however, Linden Lab is reluctant to exercise this power, choosing instead to employ a more "laissez-faire" mode of governance in Second Life (Boellstorff, 2010).

Despite their apparent dictatorship, "much of Linden Lab's governance operated at the level of setting norms, rather than managing everyday interaction," Boellstorff (2010, p. 223) observed. This sentiment was echoed more explicitly by Second Life creator and Linden Lab Chairman Philip Rosedale at an in-world town meeting about intellectual property issues in 2006 where he expressed the opinion that it would be inappropriate for Linden Lab to actively engage in resolving disputes between avatars or police Second Life. Rosedale said at the meeting: "Longer term, Second Life is going to have to develop its own law or its own standards of behavior" and encouraged Second Life users to develop "local authorities" to deal with property ownership and copyright issues (Holahan, 2006). Rosedale's sentiment was reminiscent of the earlier efforts by LambdaMOO moderators to distance themselves from resolving social conflicts among users, particularly after the outcry surrounding Mr. Bungle's cyber-rape. Following the incident, Curtis published a memo in 1993 entitled "LambdaMOO Takes A New Direction" in which he declared that the community's moderators would assume a "purely technical" role and make no decisions affecting LambdaMOO's social setting, instead implementing only consensus decisions arrived at by the community's users (Dibbell, 1993). Three years later, however, LambdaMOO's moderators issued another announcement, "LambdaMOO Takes Another Direction," in which they conceded that it was impossible for them to maintain a strictly technical presence without social ramifications (Mnookin, 1996). This 1996 memo stated:

> The line between 'technical' and 'social' is not a clear one, and never can be . . . So we now acknowledge and accept that we have unavoidably made some social decisions over the past three years, and inform you that we hold ourselves free to do so henceforth . . . The wizards will no longer refrain from taking actions that may have social implications. In three and a half years, no adequate mechanism has been found that prevents disruptive players from creating an intolerably hostile working environment for the wizards. The . . . ideal that we might somehow limit ourselves solely to technical decisions has proven to be untenable.

Despite LambdaMOO's acknowledgment in 1996 that "any technical decision may have social implications" and centralized application operators therefore must be willing to take on some responsibility for policing their online communities, several years later, Linden Lab tried to distance itself from Second Life's social disputes. "Linden does not exercise the full extent of its power over Second Life, and abstains from exercising it over the large majority of inter-avatar conduct in the world," Aiken (2008) notes. Linden Lab's forbearance in this area creates a gap in Second Life governance that is filled—at Linden's explicit encouragement and, in some cases, aided by the company's technical tools and updates—by mechanisms implemented at the other two control points: individual end-users and self-organized groups of in-world avatars.

4.4 End-User Control: Gagged & Banned

Online community-centered platforms like Second Life are in some ways intrinsically less well suited to leveraging end-users as a control point than communications applications like e-mail. E-mail facilitates primarily one-to-one interactions, between individual senders and recipients, and can therefore allow users to define bad behavior for themselves and deal with it accordingly, for instance by developing customized e-mail spam filters and block lists. This approach is less feasible in online communities, where the express purpose of the applications is to interact with a group of other users in a communal setting and some form of centralized authority, like Linden Lab, is usually needed for identifying and dealing with bad behavior. However, although community applications may rely on the involvement of a centralized authority more heavily than communications applications, there are still certain types of end-user control that can help individuals protect themselves from harassment and other forms of unwanted interaction.

These controls may allow users to customize their experience, according to their own subjective opinions about what does and doesn't constitute online misbehavior, and also helps fill the gaps in policing left by Linden. "While not ubiquitous, griefing (and the forms of conflict and misunderstanding into which griefing shaded) were too numerous to be addressed in every case by Linden Lab, a source of frustration to many residents," Boellstorff (2010, p. 195) explains, adding that "residents thus found other ways to respond to griefing." These responses ranged from simply logging off or teleporting away from griefers, to "gagging" the offending avatars or banning them from the aggrieved party's property. Gagging is essentially a mute, or ignore, command for Second Life: the command allows residents to "gag" specific other users so that they do not see anything that is said, written or done by the gagged avatar. This gagging capability dates back to many of the earlier text-based online communities, including Curtis' LambdaMOO, designed initially to imitate the various forms of avoidance people can often exercise in real-world communal settings. Curtis (1997, p. 132) explains: "It is sometimes the case on a MUD, as in real life, that one wishes to avoid getting into a conversation, either because of the particular other player involved or because of some other activity one does not wish to interrupt. In the real world, one can refrain from answering the phone, screen calls using an answering machine, or even, in copresent situations, pretend not to have heard the other party." Gagging, then, can in some sense be understood as the virtual world's answer to screening telephone calls.

In Second Life, paying users who own property can also limit access to their property to a select group of avatars or, alternatively, expressly ban specific users from entry. In these cases, avatars who are not permitted to visit a certain property encounter red "ban lines" stating "No Entry" when they approach the restricted area (Boellstorff, 2010). In this manner, users can not only ignore other residents they find particularly irritating or problematic but also avoid any interaction with them at all, at least within the confines of their own property. As banning grew in popularity, some Second Life users even began to circulate "ban lists" of known griefers, broadening the potential impact of these individual, end-user controls and simulating a "frontier ethic of taking the law into one's own hands" (Smith, 1999).

These end-user efforts—perhaps more than those made at any other control point—are particularly vulnerable to the problem of discardable identities and the ease with which individual users can easily and freely create new, additional avatars (commonly called "alts") in Second Life. Unlike Linden Lab which can, if it wants, place further constraints on the creation of these accounts by preventing banned users from registering new ones with the same credit cards, for instance, individual end-users have essentially no ability to stop their harassers from simply creating new avatars to evade their carefully constructed gags and bans. Second Lifers do have the option of only allowing a specific list of known users from entering their property—thereby banning all others, and any potential alts they might create, by default—but this measure also dilutes much of the communal nature of Second Life by limiting opportunities to interact with new, unknown residents that are valued and enjoyed by many users.

Thus, end-user controls to deal with malicious behavior in Second Life have some value in allowing individuals to define and avoid certain types of harassment but are fairly limited in their ability to scale to the larger size of the community and to deal with issues of online identity discardability. The "centralized authority" control point is better able to deal with this discardability problem, by linking avatars back to credit card accounts and implementing reputation systems, or other forms of investment in new identities, that may reduce users' incentives to continue creating unlimited new accounts, but also suffers from an inability—and in some sense, an unwillingness—to scale its efforts to deal with each individual conflict and reported abuse among the hundreds of thousands of Second Life users. To fill the gaps left by these two mechanisms, a third point of control has emerged within Second Life, and several similar online communities, consisting of self-organized groups of avatars who monitor and punish misbehavior in some semi-organized fashion, usually operating under some combination of informal self-mandate and formal recognition from the centralized, moderating authority. These emergent forms of "participatory governance," which typically require involvement on the part of both individual end-users and the centralized authority control points, are the focus of the following section.

4.5 Participatory Governance

The notion of participatory design is a common one in online games and communities. This process is broadly understood to entail a certain degree of back-and-forth between application designers and users, where a designer may construct an initial product and then modify and refine it as users engage with it, developing new uses and ideas for how it could be improved or extended. As designers respond to these new uses and user opinions—in the case of Linden Lab, by issuing new updated versions of Second Life, or implementing new functionalities like voice chat—the application's users, though they are clearly distinct from the designers, are actively involved in helping shape the community's architecture by providing ongoing feedback to the application operators. Online applications are especially well suited to this philosophy of participatory design because they can be rapidly modified, user behaviors within the applications can be readily observed by the designers, and user opinions can be easily solicited on websites and message boards or even within the applications themselves, as in the case of Second Life's avatar-circulated petitions.

"Rather than a linear, top-down process, ultimately what we find is a complex co-construction of technologies that occurs between designers, users, and the artifacts themselves," writes Taylor (2006) of participatory design schemes for online games. She continues, "Given ... that players are crucial components to the sustainability of the game, according them some power and responsibility to govern their own community and world should be a central design challenge." This extension of the participatory process and mentality not just to design but also to governance mechanisms—and indeed, the two are very much related, given that most effective governance mechanisms in the online application space require some degree of technical support from the application's design—brings us to the notion of participatory governance.

There are clearly benefits of such a participatory scheme to both individual end-users and centralized authorities like Linden Lab. To the former, it accords a greater degree of involvement and power over their own fates, mitigating (to a certain degree) some of the concerns about Linden's dictator-like authority. "[Linden Lab's] omnipotence with regard to Second Life's governance was a source of concern to many residents," (Boellstorff, 2010, p. 223) explains. "Those accused of transgressions had no way to face their accusers or appeal a decision. This led to complaints about selective enforcement." Meanwhile, Linden Lab was daunted by the prospect of policing the entirety of Second Life, especially in light of how vast its creation had grown, and "Linden Lab staff expressed a desire to delegate more of the labor of governance to residents, for instance, through the notion of covenants' on land."

Linden generally sought to limit its role in resolving internal disputes, staking out its position as responsible for the "grid stability" and technical maintenance of Second Life, but not necessarily the in-world squabbles and conflicts (Boellstorff, 2010). To deal with these social (as opposed to technical) problems, Linden Lab made some early efforts to involve residents in some familiar forms of self-regulation, modeled on real-world mechanisms. For instance, when users filed abuse reports, if Linden Lab deemed the reports worthy of further investigation or discussion it would refer the issue to a randomly chosen panel of three Second Life residents to recommend whether the offending avatar should be "executed," or expelled from Second Life. This "jury" could only offer a recommendation, rather than making the final decision, but was a notable attempt to spread the responsibility, and even some of the power, of governance to the users themselves (Aiken, 2008).

In other instances, Second Life users developed governance mechanisms on their own, independent of Linden Lab, as in the case of the Second Life Exchange Commission (SLEC), intended to regulate in-world economic markets. These self-organized bodies, however, suffered the disadvantage of having no Linden Lab-supported authority. Aiken (2008) writes, "The problem with these attempts at regulation is that they have no discernible power to punish bad actors. The peer norms are strong—citizens want businesses that will follow through on the deals made—but Linden has not delegated any enforcement mechanisms to any regulatory body." If application designers are reluctant to grant user-run regulatory bodies too much power by giving them authority over all users, it might be possible to provide users with a range of options for subjecting themselves to such user-organized regimes. Just as applications can offer users a range of identity options, they could also potentially allow users to voluntarily place themselves under the authority of any regulatory bodies organized by fellow users. Application designers could then make these decisions public, so that other users could tell which participatory regulatory authorities their peers had agreed to be governed by. Thus, voluntary adherence to user-organized regulation could serve as a signal of a user's identity and accountability, in a manner similar to the way user decisions about credit card verification of their accounts act as signals. As with end-user controls, though, user-organized governance mechanisms, whether mandatory or optional, often require the technical support of application designers and operators. Castronova (2005, p. 216) notes:

> Implementing player government is, in fact, a design decision. One cannot retain all significant power in the hands of the coding authority and then simply declare, 'Players! Make your own government!' unless one wants nothing more than a rubber-stamp body. To date, most world-builders have shied away from allowing players to form governments with real teeth. It surrenders too much control to the community of players.

Aiken (2008) is extremely critical of what he perceives as Linden's reluctance to delegate real authority to residents so they can enforce user-created norms. "In order for governance in the world to succeed more fully, Linden will either have to exercise

Table 4.1: Access Levels in English Wikipedia (Forte & Bruckman, 2008)

Administrator	Protect/unprotect pages;
	Delete/undelete pages;
	Block/unblock users;
	Special revert tools

Bureaucrat	Make administrators;
	Rename users;
	Make other bureaucrats

Steward	Change all user access levels on all Wikimedia projects
Oversight	Hide page revisions from all other user types
Checkuser	View user IP addresses
Developer	Access to MediaWiki software and Foundation servers

or distribute power," he argues, pointing out that distributing power could have the added advantages of allowing users to develop "more nuanced norms and methods for enforcing them." One of the most nuanced and well-developed models of participatory governance schemes is that of Wikipedia, the popular user-written online encyclopedia. Wikipedia's founding model was such that users would be entirely responsible for correcting and editing entries and the site's owner, Jimbo Wales, disavowed himself of any interest in monitoring and governing the project himself, early on in the site's development. However, as the website grew rapidly in size and issues like copyright infringement and falsified, biased, or merely joke entries became increasingly prevalent, users quickly identified a need for a more formal, infrastructure to flag and moderate particularly problematic entries. A hierarchy of different levels of technical access has emerged, shown in Table 4.1, involving a variety of different roles and associated powers which end-users can assume within the context of the application (Forte & Bruckman, 2008).

Linden Lab has been much less willing than the WikiMedia Foundation to delegate authority to users in the form of technical access and controls. This may be due in part to the fact that Linden Lab is a for-profit company with a clearer stake in keeping a tight grip on the development and progress of Second Life, while WikiMedia is a non-profit foundation with no vested financial interest in maintaining close control over its encyclopedia. Still, the success of Wikipedia's governance model seems to reinforce Aiken's argument that in order for such participatory schemes to be truly effective in Second Life, Linden must endow users with some concrete technical capabilities to bolster their governing authority.

4.6 Combining Control Points in Online Communities

The crucial role of Linden Lab-provided technical controls in participatory governance schemes is suggestive of the interplay and overlap between the three different points of control identified in Second Life and pseudonym-based community-centric online applications, more generally. These three control points—centralized authority (e.g. Linden Lab), end users, and communal participatory governing bodies—are by no means clearly distinguished from each other in their efforts to mitigate and manage malicious behavior. Certainly, there is a clear distinction and divide between the code-writers at Linden Lab and the users at home, operating avatars in Second Life. However, many of the Linden Lab control mechanisms—such as deciding abuse reports and attempting to set social norms via its ToS—rely at least partially on the participation and acceptance of individual users. Similarly, end-user controls like gagging and banning commands depend on the technical implementation of these capabilities by Linden coders. The lines between control points blur even more when it comes to participatory governance schemes, which require both the involvement of end-users and the technical controls developed by Linden Lab to be effective.

Each of these three control points has a role to play in Second Life when it comes to deterring and punishing misbehavior through the use of social norms reinforced by technical controls. All three control points have clear strengths and weaknesses and it is ultimately the combination of and interplay between the efforts these three points of control that offers the greatest potential for an effective governance model in Second Life. Ideally, this model would be able to balance a number of competing interests within the virtual world—the central role of pseudonyms that are clearly removed from actual world identities and the possibility of sometimes using credit card records to link some of these pseudonyms to people, the ability of individuals to decide for themselves what kinds of behavior they do or don't want to be exposed to and the need for some broader consensus on community standards and social norms, the low barriers to entry for new users in the form of free, easy-to-create accounts and the investment (of money or time) required to gain full privileges and advanced capabilities within the community. As in real-world communities, such a model is unlikely ever to satisfy all residents or solve all forms of misbehavior entirely, but a sufficiently effective and flexible system involving all three of the identified control points could go a long way towards improving and sustaining online community applications, to the benefit of both their designers and users.

Chapter 5

Facebook

> You are a bad person and everybody hates you. Have a shitty rest of your life. The world would be a better place without you.
>
> —"Josh Evans" to Megan Meier, via AOL Messenger, October 16, 2006

Twenty minutes after receiving this message, 13-year-old Megan Meier hung herself with a belt in her bedroom closet. The last thing she said in response to Evans before she committed suicide was "You're the kind of boy a girl would kill herself over." In fact, Evans was not a boy at all—he did not even exist—but Meier believed the cruel message had been sent to her by a cute 16-year-old boy who played the guitar and drums and was homeschooled and had been abandoned by his father at the age of 7. In other words, Josh Evans was more than just a screen name to Meier, he was the kind of boy a girl would kill herself over, he was someone with a fully developed personal history and identity, an identity she had become intimately acquainted with via his MySpace profile (Pokin, 2007; Steinhauer, 2008).

Launched in 2003, MySpace was one of the early successes among social networking sites, which allow users to create personal profile pages describing their interests and activities and then connect to friends and family and view their pages. In contrast to Second Life, LambdaMOO, and the other online communities discussed in chapter 4, which were explicitly designed to foster pseudonymous identities and interaction between people who did not necessarily know each other in any offline context, social networking sites like MySpace, LinkedIn, Facebook, and Google+ focus on providing users with ways to connect with their real-life acquaintances online. Most social network services encourage the creation of highly accurate (and often very detailed) online profiles that are closely tied to the users' actual identities, but they still suffer from some of the same identity verification problems as more anonymous communities since users can usually join networking sites for free, by providing nothing more than an e-mail address and some basic profile information, which may or may not be true. Josh Evans' MySpace profile, for instance, was created by Lori Drew, a 49-year-old woman whose daughter had had a falling out with Meier (Steinhauer, 2008).

Drew was subsequently identified by a girl who had helped her create the fake MySpace account and brought up on charges for her actions, unlike the large majority of griefers in Second Life and other pseudonymous communities. Indeed, just in the past decade there have been several instances of misbehavior on social network sites, ranging from cyber bullying, to trolling, identity theft, impersonation, and defamation, in which the perpetrators were successfully identified and tried in court. It would be an exaggeration to say that accountability is not a problem for social networking applications: the online disinhibition effect is evident in the numerous cases of harassment on these sites. However, when compared to online communities designed to foster interaction between strangers, social networks have had much greater success at preventing many kinds of misbehavior and holding those users who do engage in malicious activity accountable for their actions.

This chapter will examine how social networks have achieved this relative success in the domain of online accountability, looking at the identity schemes implemented by these sites and the types of misbehavior most common among their users, as well as the characteristics of the points of control and social embedding that enable the prevention and prosecution of malicious behavior. The primary focus of this chapter will be Facebook, currently the world's most popular social networking site with more than 800 million users; however, some discussion of Facebook's precursors and competitors is also included to provide a more thorough understanding of the evolution of current social network identity schemes and policies.

5.1 History of Social Network Sites

Social network sites are web applications that allow users to construct public or semi-public profiles, connect to other users, and view other users' lists of connections (Boyd & Ellison, 2008). The first such site, meeting all three of these criteria, was SixDegrees.com, launched in 1997. By 1999, SixDegrees had grown to 3 million registered users in 165 countries, but after continued financial struggles it shut down in 2001 (Penenberg, 2009). The following year, in 2002, Friendster was launched to compete with popular online dating site Match.com by matchmaking between friends-of-friends rather than complete strangers. Friendster included a list of the "most popular" users (i.e. those with the most friends listed) which incited the creation of a number of fake profiles, called "Fakesters," for celebrities and fictional characters who could collect record numbers of friends. These early fake profiles were primarily a source of entertainment rather than a means of malicious impersonation—users were aware, for the most part, that celebrities were not actually operating their supposed profiles—but Friendster made an active effort to delete Fakester profiles and even removed the "most popular" feature to deter similar forms of behavior (Boyd & Ellison, 2008).

In 2004, the deletion of Fakester accounts and widespread rumors that Friendster was considering the adoption of a fee-based model drove many users to abandon the site and join social network newcomer MySpace, launched in 2003. Beginning in 2005, MySpace was the most visited social network site in the world, with 100 million users by August 2006 and an estimated value of $12 billion in 2007. Then, on April 19, 2008, MySpace was overtaken by Facebook in monthly global unique visitors, though it would continue to lag behind in U.S. users for another year.

Launched in February 2004 by CEO Mark Zuckerberg as a site exclusively for users with Harvard.edu e-mail addresses, Facebook gradually expanded to other universities and, finally, to the wider population. Today, with more than $4 billion in annual revenue and 800 million user accounts it continues to be the most popular social network site in the world. One of Facebook's newest competitors Google+, the social networking service launched by Internet search giant Google in June 2011, has come under fire recently months for its aggressively enforced "real names policy" which requires users to identify themselves with their full first and last names. This controversial policy is similar to the requirements set forth in Facebook's (typically less stringently enforced) Terms of Service agreement and gets at a crucial characteristic of online identities in popular social network applications: how closely tied they are to users' real-life identities.

5.2 Real Name Policies & Social Network Identity Schemes

In January 2011, Facebook shut down the profile of Chinese activist and blogger Zhao Jing because he had opened his account under his professional pen name, Michael Anti. Later that year, Google+ suspended the accounts of several users, such as sex blogger Violet Blue and engineer Limor "Ladyada" Fried, who had signed up to use the service under nicknames or pseudonyms (McCracken, 2011). These users had violated the Terms of Service (ToS) they agreed to when they joined the social network sites, Facebook and Google explained to outraged members. Both sites' ToS stipulate that users list their full, real names on their profiles, a requirement that upset many, leading to the so-called "nymwars" over members' rights to use pseudonyms on these sites and the creation of the website "My Name Is Me" (my.nameis.me) dedicated to convincing "online services—including social networks such as Facebook and Google+—to allow users to identify themselves by whatever name they choose."

Thinking back to the troubles Second Life and other pseudonymous communities encountered due to their users' relative anonymity, it's not difficult to imagine why social network sites might hope to avoid similar forms of misbehavior by forcing users to associate their profiles with their real names. "The Internet would be better if we had an accurate notion that you were a real person as opposed to a dog, or a fake person, or a spammer," Google executive chairman Eric Schmidt said at a media conference in August 2011, defending Google+'s real names policy. "If we knew that it was a real person, then we could sort of hold them accountable, we could check them," he added (Pfanner, 2011). Critics contend that real name policies are damaging to a wide range of users who may desire anonymity (or pseudonymity), including people who regularly employ professional pseudonyms, at-risk populations such as abuse survivors, LGBT teens, and political dissidents, and users who want to maintain separate personal and professional identities. But these arguments have held little sway with the most prominent social network organizations that largely maintain the accountability provided by mandating real names outweighs the privacy and free speech benefits afforded by anonymity. Facebook marketing director Randi Zuckerberg made a similar statement the month before at a panel discussion, telling the audience, "I think anonymity on the Internet has to go away ... People behave a lot better when they have their real names down" (Bosker, 2011).

But real name policies are not the key to social network sites' relative success at holding users accountable for their behavior. Indeed, in many cases such policies are largely unenforced and unenforceable. Let's return for a moment to Lori Drew's decision to set up a MySpace account for Josh Evans—obviously not her real name. MySpace, at the time, had an early variant of a real names policy included in its Terms of Service, which stipulated that members must submit "truthful and accurate" registration information (Stelter, 2008). MySpace, like Facebook and Google+, technically requires members to submit their true names but did not demand any verification from new users upon registration. Even now, it is possible to create profiles on most social networks by submitting nothing more than an e-mail address and self-reported profile information that the controlling company has no means of easily identifying as false unless other users report it to be inaccurate. A quick search on Facebook for users with the name "Santa Claus" or "Tooth Fairy" rapidly reveals that the legacy of Friendster's Fakester profiles lives on in today's social networks.

It is true that Facebook and Google+ have both shut down accounts for using false names, but the vast majority of users who wish to maintain their profiles under nicknames or patently fake ones continue to do so (Boyd, 2011). In some cases, like that of Lori Drew and Josh Evans, this may be because the operating site has no way of knowing that the name is false. In others, such as the hundreds of Santa Claus Facebook profiles, presumably, it is because the operating site does not really care about forcing everyone to use their real names. Google+ even adjusted its policy earlier this year to allow people to use nicknames and pseudonyms on the site, if they can "prove to Google that [they are] known by that name elsewhere, in published material or on other social networks" (Miller, 2012). This flexibility stems from the fact that allowing users to maintain profiles under names like Santa Claus—or even Josh Evans—does not actually hinder the site's core accountability mechanism, which runs much deeper than the real name policy. Social networks rely on the deep embedding of their applications within users' actual, real-life social spheres to promote accountability. This embedding occurs regardless of whether a member uses a real name or not and it is central to the very function of social network sites, that they, by design, connect users with people they know in real-life. In other words, these online connections are anchored in offline relationships and it is these "anchored relationships" (Zhao, Grasmuck, & Martin, 2008) that social networks' accountability mechanisms rely upon, not a preponderance of real names or the policies mandating them. That's a somewhat subtle distinction but an important one: the name on a given social network profile does not dictate how accountable the owner is (and thinks she is) for her actions, the other people she is connected to in that network do.

Think back to Lori Drew; she violated MySpace's Terms of Service and operated an account under a completely fictional name—but she was still identified and tried for her actions in court. The absence of her real name, in other words, was not a significant obstacle to holding her accountable. In Drew's case, as in many instances of malicious behavior on social networks sites, this was because her actions directly concerned and were deeply embedded within her actual, real-life community. She harassed a girl who was a former friend of her daughter's, she did so together with another teenager who knew the involved parties—because Drew's online actions were so deeply anchored in her offline social sphere, it turned out to be relatively straightforward to identify her and hold her accountable in court.

5.3 Self-Presentation on Social Network Sites

Just because social network site profiles tend to be anchored in real-life relationships does not mean they are necessarily entirely accurate, several researchers have noted. Where users of more anonymous online community sites such as Second Life may create avatars that allow them to assume alternate or unexpressed elements of their identity, social network site users may select specific information and photographs for their profile to construct a somewhat idealized version of themselves (Zhao et al., 2008). Anonymous online communities may permit users to completely reinvent themselves, but even "nonymous" communities like Facebook and Google Plus allow for a certain degree of identity performance and carefully constructed self-presentation (Douglas & McGarty, 2001). On social network sites, identity performance has been found to be closely related to patterns of self-presentation on Internet dating sites where many users create profiles that reflect "an ideal as opposed to actual self" as they struggled with trying to remain truthful while still projection "a version of self that was attractive, successful, and desirable" (Ellison, Heino, & Gibbs, 2006) whether consciously or not. In one study of Internet dating participants, 94 percent of respondents said they had not intentionally misrepresented themselves in their online profiles and 87 percent said they did not think that such misrepresentation was acceptable, but 86 percent said that they had encountered other users who misrepresented their physical appearance (other commonly misrepresented features included relationship goals, age, income and marital status, the researchers found). "Misrepresentation was not always intentional and occurred in three ways: through representation of an inaccurate self-concept, fudging demographic information such as age . . . and portrayal of an idealized or potential future version of the self," the research team noted (Gibbs, Ellison, & Heino, 2006).

Internet dating participants are, of course, constrained in how much their online profiles can differ from their real-life identities by the prospect of face-to-face encounters with the people they meet online. Thus, online daters must balance their desire to idealize their self-portrayal with the risks that if their profile strays too far from the truth these discrepancies may be revealed when they finally reach the stage of in-person dating (Gibbs et al., 2006). On social network sites, the constraints on manipulating one's self-presentation are even tighter, since many users have already had face-to-face encounters with the majority of other users they are connected to via online social networks. Still, given that early social networks like Friendster emerged from the online dating trend, it is not entirely surprising that social network identities would reflect some of the same (potentially unintentional or subconscious) self-idealization and careful construction of profiles to project traits like popularity and attractiveness by making public one's list of friends or particularly flattering photographs (Zhao et al., 2008).

In a survey of first-year Michigan State University students, Lampe, Ellison, and Steinfield (2006) found that students overwhelmingly reported that their Facebook portrayals described them accurately. The results also showed that Facebook seemed to be most commonly used by students to maintain previous relationships and investigate people they'd met offline, rather than searching for strangers online with the intention of then initiating offline contact, as traditionally occurs on dating sites. Thus, social network profiles are self-constructed and often slightly idealized versions of users but their anchoring in real-life social relationships typically limits the extent to which these identities can be falsified or exaggerated.

5.4 Misbehavior & Points of Control in Facebook

The anchoring of social network identities in the real world provides some accountability, but it does not limit users' ability to misuse social network applications for malicious purposes, though this misbehavior often mimics the nature of social network sites in that it, too, is often anchored in a real-life hostility or vendetta, as in the case of Lori Drew.

5.4.1 *Applause Store Productions Ltd. and Matthew Firsht v. Grant Raphael*

British cameraman Simon Firsht was browsing Facebook on July 4, 2007, when he discovered a profile for his twin brother, Mathew Firsht, but something about the page seemed wrong to Simon. For one thing, the profile seemed to falsely indicate his brother's political affiliation and sexual orientation (he was a member of groups titled "Gay in the Wood ... Borehamwood" and "Gay Jews in London"). Additionally, Mathew's Facebook page was associated with a company profile called "Has Mathew Firsht lied to you?" which alleged that that Mathew owed large sums of money and regularly made excuses to avoid paying his debts (Bingham, 2008). Mathew confirmed to his brother that he had not set up the profile and they contacted Facebook, which promptly shut down the account in question. The brothers' then went to court and obtained a search order which required Facebook to disclose the e-mail address associated with the offending profile as well as the IP address of the computer that was used to create it on June 19, 2007.

Firsht's lawyers quickly traced the IP address to the flat of Grant Raphael, who had been a good friend of Mathew's at school in Brighton and worked with him at the TV production company Power House until 2000, when Firsht resigned from the company on Raphael's encouragement but later found out that Raphael had taken over his office immediately following the resignation. The two stopped speaking to each other shortly thereafter. Raphael denied creating the false profile for Firsht, claiming that it must have been someone else using his home computer. In the 2008 case Applause Store Productions Ltd. and Matthew Firsht v. Grant Raphael, tried in the British High Court, Deputy Judge Richard Parkes QC dismissed this defense as "built on lies" and awarded Firsht (and his company Applause Store Productions Ltd.) 22,000 in damages for libel and breach of privacy.

In his decision (*Applause Store Productions Ltd and Matthew Firsht v Grant Raphael*, 2008), Judge Parkes writes:

> I found the Defendant's explanation for the Facebook usage on 19th June utterly implausible from start to finish. The proposition is that on 19th June a complete and random stranger, visiting the Defendant's small flat for the first time, should first have gone into the Defendant's study and started using his computer, without permission, over a period of about an hour, without being observed, should then have created a false and hurtful Facebook profile about a man whom the Defendant knew well and had fallen out with, containing private information and other information which few people apart from the Defendant would have known, and should have searched from that profile for a number of people known to the Defendant. In my judgment, the proposition has only to be stated to be rejected as utterly far-fetched. ... Moreover, in my judgment there was a degree of needle, to put it no higher, on the Defendant's part. He had been rejected by his old boyhood friend several years before, and the old friend had prospered greatly in the intervening years, while the Defendant had not. He had a motive (if not a justification) to inflict some damage on Mr Firsht.

There are several elements of this decision worth noting that deal with issues of online accountability and attribution. The first is the Judge's decision to extend attribution based on an IP address from a physical machine to a specific person. In his decision, Parkes even notes that there is "no primary witness evidence that the Defendant was responsible for the creation of the false Facebook pages." Instead, the Judge relies on the unlikelihood of Raphael's proposed alternative scenario (a stranger creating the profile from Raphael's home computer) and the two men's animosity to conclude that Raphael is guilty. Once again, the embedding of the online social network application within a real-life social relationship is absolutely crucial to determining who will be held accountable. Like Lori Drew, Grant Raphael used a fake social network profile to exact revenge on someone with whom he had a real-world relationship. And like Lori Drew, he was identified and held accountable in court.

5.4.2 *R v. Hampson*

Harassing, bullying, impersonating and defaming real-life acquaintances are common forms of malicious behavior on social networks sites, but some users also take advantage of these applications to harass complete strangers, a practice known as "trolling" which bears close resemblance to the griefing activities observed in anonymous communities like Second Life and World of Warcraft. Trolls typically post derogatory or insulting comments and images on social network profiles and pages for people they have no real-life connection to, especially memorial or tribute pages for the dead. For instance, on February 15, 2010, Bradley Paul Hampson, a 29-year-old autistic Australian, used his Facebook account (created under the pseudonym Dale Angerer) to post on the tribute page of recently-murdered 12-year-old Elliott Fletcher. Unlike the majority of other visitors to the page, who expressed condolences to Fletcher's family and words of sympathy for the stabbed schoolchild, however, Hampson posted a picture of Fletcher's face with the words "WOOT IM DEAD" superimposed on the image. Later that day, Hampson posted another picture on Fletcher's tribute page, this one depicting Fletcher's head in the hopper of a wood-chipper, with blood gushing out of the chipper and the profile picture of "Dale Angerer" standing next to it, with a caption bubble reading: "Hi, Dale Angerer here I fully endorse this product. This woodchipper can mince up any dead corpse or your money back guarantee." Later that month, Hampson also posted similarly mocking and offensive images on the tribute page for 8-year-old Trinity Bates, who had been kidnapped and found dead in a storm water drain near her house (*R v. Hampson. QCA 132; BC201104379*, 2011).

There were multiple trolls posting on Fletcher's and Bates' tribute pages, but when the Australian police decided to investigate these offensive posts they only identified one user as being within their jurisdiction: Hampson. Though Hampson had joined Facebook under a pseudonym, Dale Angerer's posts were traced back to the IP address of his home computer. The police then seized this computer and discovered the files of the images posted by Angerer (including "child exploitation material") containing the images that had been posted by Angerer. Hampson was taken to court on charges of distributing child exploitation material and "using a carriage service to menace, harass or cause offence." He pleaded guilty and was sentenced in March 2011 to serve a three-year jail term, though his sentence was later shortened to six months on appeal (Bentley, 2011).

Using a carriage service to menace, harass or cause offense to others is a crime carrying a maximum penalty of three years improvement under Australia's Criminal Code and "although [it is] not a trivial offence, [it] is not one which the legislature regards with great severity," according to the Hampson ruling. The Supreme Court of Queensland, in its June 2011 decision shortening Hampson's sentence, also noted the relative gravity of trolling with respect to other Internet offenses, writing:

> The use of the Internet to harass and bully to the extent that the victim suffered lasting psychological harm or was driven to suicide may be thought to be a more serious category of offending. So too, would be the use of the Internet to publish false and defamatory matter leading to the loss of the victim's good reputation and/or the collapse of a business. This is not to say that the subject conduct was not extremely serious. It was ghoulish and disgusting by any reasonable standards and its inevitable consequence was to cause emotional pain and distress to grieving relatives and friends of the deceased children.

In other words, in the eyes of the Supreme Court of Queensland at least, Hampson's actions were less serious than Lori Drew's or Grant Raphael's misuse of social networking sites (ironically, of the three, Drew is the only one who was acquitted). All three, however, were held accountable in court, reinforcing the idea that accountability mechanisms for social network applications, which are so deeply anchored in relationships based in the physical world, closely mimic the accountability schemes we see in the physical world, unlike the mechanisms found in more anonymous virtual communities like Second Life and Wikipedia.

Drew, Raphael, and Hampson are not the only people who have been taken to court for their behavior on social network sites but, undoubtedly, there are many Facebook trolls, defamers, and bullies who have never been put on trial. Their three cases highlight three common categories of misbehavior on social networks sites—cyberbullying, defamation, and trolling—as well as three crucial points of control involved in trying to mitigate that behavior: the companies that own the social network sites, the individual users, and the courts. The following sections of this chapter will look at each of these control points in greater detail, examining the mechanisms available to each of these three actors for mitigating misbehavior and holding malicious users accountable for their actions, as well as the way these different control points can coordinate and combine their powers to achieve more effective means of enforcing accountability standards.

5.5 Centralized Authority: Operating & Owning Companies

Companies like Facebook, Google, and MySpace that own and operate social network applications are a natural point of control when it comes to defining, deterring and punishing malicious activity on these sites. These companies, like Linden Lab and those that operate anonymous online communities, can generally exercise this control at several different levels: by dictating the site's terms of service agreement, by shutting down accounts found to be violating these terms of service, by investigating user complaints and abuse reports, and by tracing activity on their sites back to specific IP addresses for purposes of attribution.

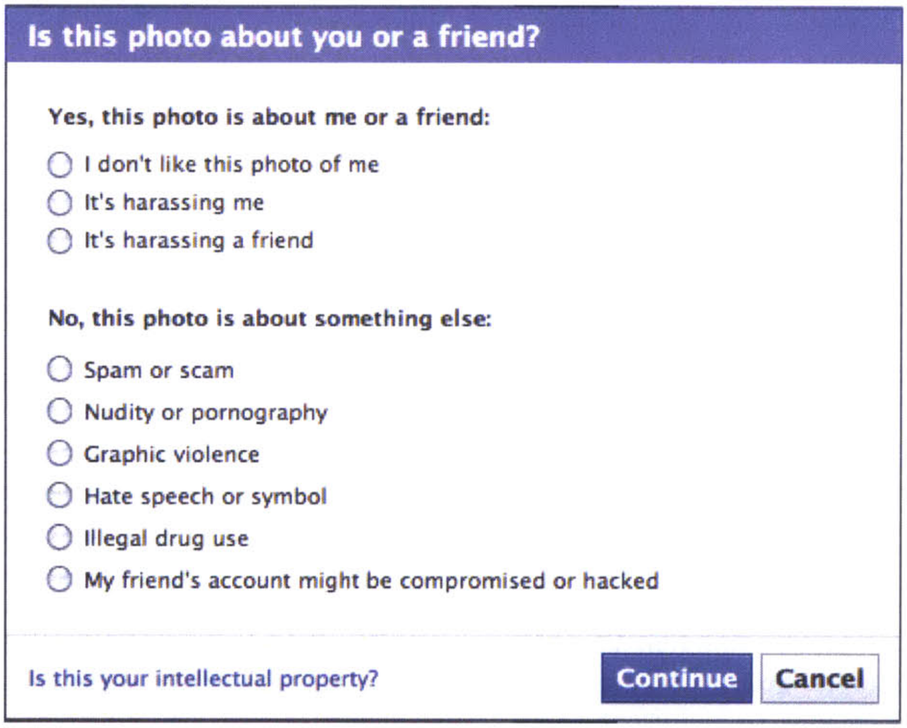
Like many other online applications, social network sites are typically governed by terms of service agreements which dictate what constitutes appropriate and inappropriate use of the application. In some instances, such agreements have even been argued to be legally binding under United States law. For instance, the charges brought against Lori Drew were based on her violations of MySpace's Terms of Service (ToS) and the idea that by breaching this agreement (which required she use her real name on any profile she created), Drew had accessed MySpace "in excess of authorized use" and therefore violated the Computer Fraud and Abuse Act (CFAA). Drew was acquitted but the deciding Judge did not rule that ToS violations were beyond the scope of the CFAA, merely that in Drew's particular case the agreement was too vague since it was unclear whether any or all violations of terms of service rendered the access unauthorized (and therefore criminal). In fact, District Judge George Wu explicitly states in his 2009 decision of the case that "an intentional breach of the [MySpace ToS] can potentially constitute accessing the MySpace computer/server without authorization and/or in excess of authorization" (*U.S. v. Lori Drew*, 2009).

The question of whether Terms of Service violations constitute criminal activity under the CFAA continues to be an area of contention in U.S. case law, but whether or not they carry the weight of law in some jurisdictions, ToS agreements and registration requirements can be immensely influential in establishing the social norms of a social network site. When a company like Facebook or Google writes Terms of Service for a social network application, it can, to a large degree, dictate the general level of identifiability of its users, especially when coupled with its ability to require users to provide certain information (e.g. name, e-mail address, etc.) The company-instituted Terms of Service and registration requirements are not by any means a guarantee that all users will abide by that agreement or provide accurate information upon registering, but they do set a baseline norm for the given application and these norms can vary greatly, depending on the preference of the governing company. Gross and Acquisti (2005) explain:

> The pretense of identifiability changes across different types of sites. The use of real names to (re)present an account profile to the rest of the online community may be encouraged (through technical specifications, registration requirements, or social norms) in college websites like the Facebook, that aspire to connect participants' profiles to their public identities. The use of real names may be tolerated but filtered in dating/connecting sites like Friendster, that create a thin shield of weak pseudonymity between the public identity of a person and her online persona by making only the first name of a participant visible to others, and not her last name. Or, the use of real names and personal contact information could be openly discouraged, as in pseudonymous based dating websites like Match.com, that attempt to protect the public identity of a person by making its linkage to the online persona more difficult.

Notably, in all of these cases the different degrees of identifiability are in large part dictated by the owning and operating companies through their ToS, user requirements, and technical specifications.

Writing Terms of Service agreements is one thing, enforcing them, of course, is quite another. and often requires considerable time and energy. At Drew's trial, Jae Sung, then the Vice President of Customer, testified that there was no possible way for MySpace to determine how many of the 400 million MySpace accounts violated the company's Terms of Service and many companies have tended to wait for other users to report any violations rather than trying to seek them out. Most social network sites, including Facebook and Google Plus, have a feature where users can file "abuse reports" if they experience any harassment or misuse of the application though, unlike in Second Life, these reports are generally handled entirely at the discretion of the company rather than being referred to a jury of users. In 2009, responding to a slew of cyberbullying cases, Facebook unveiled a redesigned abuse report template which provides "more granular reporting categories" and additional fields to "detail the location of abuse that occurs in videos or text." Facebook includes "Report" links next to every photo, video, or note posted on the site, but emphasizes on its blog the important role individual users play in helping the company monitor harmful content. A blog post announcing the new abuse reports states: "The information you provide helps our international team of professional reviewers prioritize reports and know what they're looking for when reviewing the content ... We rely on you to let us know when you see objectionable content" (Ghastin, 2009). These abuse reports, shown in Figures 5-1, 5-2, and 5-3, offer Facebook users the opportunity to specify their concerns and offer them a choice between contacting the offending user directly or reporting the content to Facebook, depending on which control point they wish to appeal to—their fellow end-users or Facebook.



Is this photo about you or a friend?

Yes, this photo is about me or a friend:

○ I don't like this photo of me
○ It's harassing me
○ It's harassing a friend

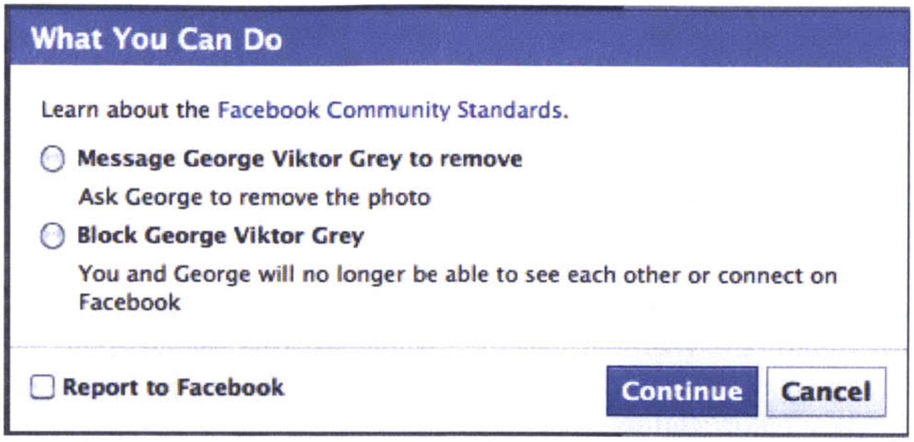
No, this photo is about something else:

○ Spam or scam
○ Nudity or pornography
○ Graphic violence
○ Hate speech or symbol
○ Illegal drug use
○ My friend's account might be compromised or hacked

Is this your intellectual property?

Figure 5-1: The first step of reporting a photo on Facebook involves designating the reason for the report.

Over the course of the past decade, as social network site misuse has attracted more media attention and legislative efforts, some companies have made more active efforts to identify and shut down false accounts. In a 2005 interview, Mark Zuckerberg noted that Facebook uses algorithms to help employees analyze how "real" users were and identify fake accounts (Raynes-Goldie, 2010). More recently, Google has come under fire for its active policing of Google+ accounts that use pseudonyms or nicknames (McCracken, 2011). After they identify profiles containing false or malicious information, companies like Facebook and Google have relatively few options for holding these users accountable. Essentially, the operating companies can suspend or terminate these accounts but have generally not employed any of the more



What You Can Do

Learn about the Facebook Community Standards.

○ **Message George Viktor Grey to remove**

Ask George to remove the photo

○ **Block George Viktor Grey**

You and George will no longer be able to see each other or connect on Facebook

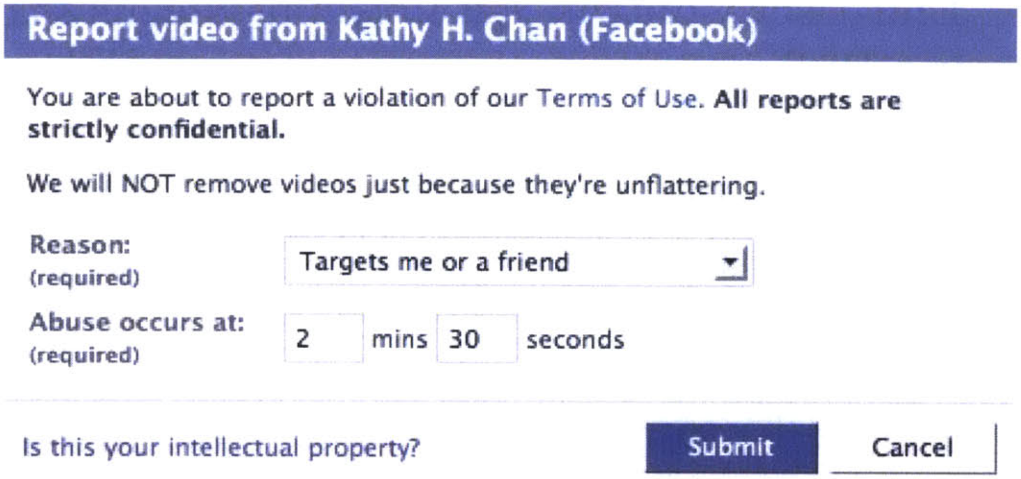
□ **Report to Facebook**

Figure 5-2: The second step of reporting a photo on Facebook allows users to choose between sending a message to the offending user, blocking that user, or reporting the content to Facebook.

"reputation-based" strategies seen in communities like Second Life and Yelp. Generally speaking, social network profiles are either active or shut down—there are few intermediate status levels that users can be demoted to, or extra privileges that can be taken away.

This reliance on account termination as a means for holding users accountable on social network applications may be due to the fact that shutting down a user's social network profile is in many ways a much more effective punishment than, say, deleting someone's Second Life avatar or World of Warcraft character, because it is not as easy to immediately create a new, different online identity for the same application. Certainly, a user kicked off Facebook can immediately create a new, free account using a different e-mail address, but the likelihood of that user wanting to create a completely different identity on a social network site and use it to connect with a different group of people than before is quite low, since most people use these applications to represent their actual identities and connect to their actual-world friends. Thus, the prospect of creating an unlimited number of free, new social network identities is of limited appeal.

Furthermore, building up a robust social network profile requires a considerable investment of time and energy on the part of the user to identify and add friends, join communities, post photos, and other related activities. Reputation-based online identities for applications like Yelp, Second Life, and World of Warcraft may also require an investment of time and energy but social network profiles are unlike these other forms of online identity in that they can really only be effectively constructed by the user they belong to. In other words, it is possible in many online communities to purchase reputational identity elements by paying other people to devote the time necessary to build up reputations and acquire certain tools and privileges (in the gaming world, this practice is commonly known as gold farming, in which players purchase in-game currency, or gold, for real money from other players). By contrast, richer users cannot easily pay someone else to build their Facebook profiles



Report video from Kathy H. Chan (Facebook)

You are about to report a violation of our Terms of Use. **All reports are strictly confidential.**

We will NOT remove videos just because they're unflattering.

Reason: (required) Targets me or a friend

Abuse occurs at: (required) 2 mins 30 seconds

Is this your intellectual property?

Figure 5-3: Reporting a video on Facebook also requires users to specify the reason and specific time at which the alleged abuse occurs.

when that process involves intimate knowledge of their social spheres, interests and activities. Thus, social network identities are less easily discardable than those of many other online applications and this allows companies like Facebook and Google to use account termination as a relatively effective means of punishing and deterring unwanted behavior.

Finally, as a last recourse in most circumstances, companies that own social network sites can trace user activity to specific IP addresses and retrieve logs of user activity from those addresses. This tracing, which is generally only performed at the request of a court order, was the means by which Mathew Firsht identified his defamer as Grant Raphael and also the way the Australian police identified Bradley Paul Hampson as Dale Angerer. Facebook cannot, of course, actually identify who is performing the activity it records, only the account name and IP address associated with that activity. This could be considered an even weaker form of attribution than, for instance, Linden Lab's ability to trace many of its users back to a registered credit card. Credit cards at least are usually traceable to an individual, while IP addresses may be spoofed, or may correspond to public areas, such as coffeeshops, or, even in the case where they do lead back to a specific machine indicate nothing about who precisely was using that machine at the time. Hampson confessed to his trolling activities, rendering this point irrelevant, but Raphael based his entire (unsuccessful) defense on it, claiming that someone else had used his computer to defame Firsht. Whether or not Raphael's claim was true (the court clearly thought it wasn't), it is clearly not implausible to imagine someone using public Wi-Fi or a computer belonging to someone else to carry out some malicious action on a social network site and, Raphael's case notwithstanding, this possibility places severe limitations on the usefulness and reliability of company-maintained IP address logs as a tool for accountability on social network sites.

5.6 End-User Control: Establishing Norms & Blocking Bullies

With such limited ability to reliably trace users, companies like Facebook depend to a very great extent on their users to prevent misbehavior among themselves, both by reporting malicious activity to the company and by establishing social norms for the online community that mimic those of the physical world. Zhuo (2010), Facebook's Manager for Product Design, writes that the company's approach is "to try to replicate real-world social norms by emphasizing the human qualities of conversation." Facebook does this by placing user's profile photos, names, and sometimes other biographical details such as hometown, alongside their public comments to "establish a baseline of responsibility." In other words, Facebook's technical design is intended to help create social norms that rely on widespread adoption and enforcement by individual users.

Facebook users have some technical tools at their disposal to avoid people they find particularly unpleasant on social network sites. For instance, as in Second Life, users can choose to block another Facebook account from viewing their profile or contacting them in any fashion through the application. However, in cases of cyberbullying and trolling, or other instances where a user is being insulted and harassed in a public forum visible to many other Facebook users, these individual controls may be inadequate. Under these circumstances, users have few options beyond reporting their concerns to Facebook (or, in some limited cases, appealing to the legal system). Notably, there are very few effective forms of "participatory governance" on social network sites. Where members of online communities like Second Life and Wikipedia have established some intermediate regulatory structures within these communities, social network sites are largely devoid of any self-organized user-based governing bodies. Users can—and do—join groups on Facebook protesting changes in the site's privacy settings or layout, but this relatively weak form of coordination within the online community itself is generally ignored by the company and, more often than not, quickly abandoned by participating users.

Though Facebook and other social network sites have been reluctant to empower user-organized regulatory and reputation schemes, they have generally been viewed as reasonably responsive to user complaints. Unlike Linden Lab, which explicitly stated its wish not to be involved in in-world user disputes, Facebook plays a much more active role in moderating conflicts on the site. For instance, when Mathew Firsht reported the defamatory profile made under his name, Facebook promptly deleted it. Still, despite their active involvement in resolving user disputes, Facebook has been criticized by those who feel it provides inadequate privacy controls to individual users, especially those which would help members manage multiple versions of their profile for different groups of people including colleagues from work, friends from school, family members, and others (DiMicco & Millen, 2007). The design of Google+, which allows users to put their friends into different "circles" and designate very specifically what members of each circle can and can't view on their profile, was intended in part to respond to precisely these concerns.

Ultimately, though, even if Facebook provides relatively few tools to end-users to help them organize and regulate each other's behavior, the atmosphere of accountability present in social networking applications like Facebook is due primarily to the social norms established and upheld by these users. Zhuo (2010) explains:

> Facebook ... encourages you to share your comments with your friends. Though you're free to opt out, the knowledge that what you say may be seen by the people you know is a big deterrent to trollish behavior. This kind of social pressure works because, at the end of the day, most trolls wouldn't have the gall to say to another person's face half the things they anonymously post on the Internet.

On social network sites, as in the real world, however, even strong social norms are not a sufficient deterrent to eliminate all harassment or malicious activity and when users encounter such behavior they may choose to communicate their concerns to their fellow end-users, report them to the company that owns the site, or take their grievances to the courts.

5.7 The Role of Legislation & Law Enforcement

The courts are at once the most powerful and least powerful point of control when it comes to accountability mechanisms for social network sites. They are the most powerful because they are the only body with the authority to impose penalties like fines and imprisonment on actual people, rather than their online identities. Facebook, for all its means of holding users accountable, cannot extend its disciplinary reach beyond the profiles on its site to the actors who created them. But at the same time, courts are tremendously limited in their ability to impose these penalties by issues of jurisdiction. Facebook's users span the entire globe, with less than one quarter reporting that they live in the United States.

Thus, even as several states move forward with legislation to restrict cyberbullying, the reach of such statutes is decidedly limited when viewed in the greater global context of applications like Facebook. However, cases like Drew's and Raphael's indicate that while the scope of such legislation may be narrow it is not entirely irrelevant, given how much malicious activity on social network sites is directed towards people the offending user knows in real life and, in many cases, the perpetrator therefore resides within the same jurisdiction as the victim, allowing for the latter to take legal action. Hampson's case, by contrast, illustrates precisely how difficult it is to use purely legal means to rid sites like Facebook of trolls when Hampson's was the only account, of many which posted similarly offensive content, that the Australian police were able to identify as within their jurisdiction.

Despite its limitations, cases like those of Drew, Raphael, and Hampson, clearly indicate that laws and lawsuits can play a non-trivial role in holding social network users accountable for certain types of misbehavior, specifically those which violate the legal statutes within the user's jurisdiction. It is worth noting, however, that the role of the courts is deeply tied to and dependent on the involvement of the other two points of control discussed above: the owning company and the individual users. Each of the three cases discussed in this chapter illustrates that dependence in a different way. In Drew's case, it was necessary for another MySpace user to come forward and identify Drew as the creator of the profile for Josh Evans in order for her to be brought to trial. For Raphael to be found guilty of defamation, the court relied on IP address and activity logs provided by Facebook, and identifying Hampson as the troll Dale Angerer required obtaining similar logs from Facebook though, since Hampson pleaded guilty, the charges rested less entirely on the Facebook-provided data.

Since social network sites are intended in many ways to mimic the social accountability structures of the physical world, it is perhaps not surprising that the legal enforcement structures have become more concerned with such sites than they have with many other online communities. Second Life griefers, after all, have rarely (if ever) found themselves in court even though their actions are roughly comparable to those of Facebook trolls like Hampson. In part, this discrepancy may stem from the greater anonymity of communities like Second Life, which makes the direct attribution needed for a court case more difficult to achieve. However, in some sense, Linden Lab possesses as much information, if not more, about its customers than Facebook since Linden can also trace account activity to specific IP addresses and, furthermore, collects credit card information for many of its users. An alternative explanation might be that the types of misbehavior witnessed in Second Life are generally regarded as less serious than the cyberbullying and harassment seen on Facebook. But this distinction is rather arbitrary and largely unfounded since, excepting cases like Meier's where one user's actions appeared to cause a suicide, the consequences of much of the harassment and trolling on Facebook seem fairly comparable to those of the griefing activities in Second Life. Fundamentally, it seems, there may be something about the inherent function and structure of social network sites and the ways in which these virtual worlds so closely mimic social elements of the physical world that drives users to try to apply to the virtual domain one of the primary accountability mechanisms available to them in the physical world: the rule of law.

Because social network sites do, in fact, belong to the virtual rather than the physical world, there are a number of ways social network members can protect their anonymity and increase their chances of avoiding legal charges—whether by masking their IP address, using public computers to access the sites, or keeping their actions secret from others. It is possible that current members wishing to engage in malicious activity on social network sites will learn from the mistakes of users like Drew, Raphael, and Hampson, rendering legal recourse an increasingly ineffective accountability mechanism in this area. However, it is perhaps equally likely that social network sites and their users will instead move in the direction of an even deeper embedding within real-life social relationships, as their services and user bases continue to develop and expand, possibly making legal regulations and enforcements even more applicable and relevant.

5.8 Identity Aggregation & the Future of Accountability on Social Networks

The continued growth and popularity of social networking sites indicate that they will play an important role in shaping the future of the Internet, but it remains uncertain what the future of accountability mechanisms on these sites will look like. Examining the cases of Drew, Raphael, and Hampson, in turn, can lead to very different conclusions about what direction social network accountability mechanisms are headed in. While the decision in *USA v. Lori Drew* implies that company-drafted Terms of Service agreements, supported by the force of law, may be the crucial element in enforcing behavioral norms, the case of *Applause Store Productions Ltd and Matthew Firsht v. Grant Raphael* suggests that IP addresses and user activity logs may instead be the key to holding users accountable for their actions on social network sites. By contrast, the case of *R v. Hampson*, in which only one of many trolls is held accountable for his actions, implies that the impact of lawsuits in this realm may be becoming negligible.

This chapter has discussed three primary points of control on social network sites. First, there are the owning and operating companies, which can dictate terms of service, shut down user accounts, and trace all activity back to specific accounts and IP addresses. Second, individual end-users can establish and maintain the virtual communities' social norms, block other members they find particularly offensive and, in more extreme cases, report these users to the controlling company or bring them to court. Finally, legal and legislative bodies can enact and enforce laws pertaining to the abuse of these social network applications, so long as the actual identities of the offending users can be satisfactorily established. Importantly, the success of the accountability mechanisms employed by any one of these three control points relies heavily on the involvement of the others. Facebook can design its site to encourage certain social behaviors and norms but their enforcement will depend on the end-users. Those same users can block and report other members but only if provided with the technical tools to do so by Facebook. Similarly, lawsuits concerning social network misbehavior depend on individual users filing and testifying at them, as well as, in many cases, IP logs provided by the site's owner.

This interdependence suggests that all three points of control may continue to operate, at least in some fashion, into the future but the interplay between them and relative power are likely to shift as social network sites continue to evolve. One trend that does seem to be emerging in recent years is the aggregation of multiple online identities into single social network-based profiles. More and more, through services like Facebook Connect and Google accounts, individuals are using the identities they've created on social network sites to access and comment on other types of online applications, ranging from newspaper and entertainment sites to online merchants. For users, this sort of account aggregation, discussed in greater detail in chapter 9, can have the benefit of greater convenience—fewer passwords and user-names to remember—and also allow them to view items, comments, or articles that have been purchased, written, or flagged by their friends. Some have also praised the way in which extending social network profiles to other applications has seemed to increase the accountability of, and therefore improve behavior on, these other types of websites, as well. Or, as McCracken (2011) put it: "When Silicon Valley blog TechCrunch dumped its commenting system for one that required readers to sign in with their Facebook credentials, the IQ of the average poster seemed to instantly jump by about 50 points."

Some researchers have predicted that the links between social network profiles could be leveraged to even greater effect to improve the social norms and behavior of a wider variety of Internet applications. For instance, Donath and Boyd (2004, p. 81) write:

> It is possible to imagine a scenario in which social networking software plays an increasingly important role in our lives. For instance, e-mail is becoming increasingly unusable as spam fills inboxes, keeping one step ahead of the filtering heuristics. Perhaps a social network based filter is the solution—e-mail from your connections would always go through, and perhaps from the next degree out. Anyone else would need to go through a chain of connections to reach your inbox (or at least, to reach it with the seal of approved non-junk).

Aggregating online identities in this manner does have costs to end users, most notably it can make it increasingly difficult for users to maintain a diverse set of unrelated online identities to correspond with the diverse set of Internet applications available. This can be particularly problematic for users who wish to maintain and manage "multiple online presentations of self" (DiMicco & Millen, 2007). Concerns about privacy notwithstanding, though, this trend towards identity integration through social network profiles hints at the dominating role these identities may have online in years to come and the crucial importance and increasingly widespread applicability and relevance of the accountability mechanisms associated with them. If social network identities are to become one of the dominant identity schemes of the Internet, crossing over many different applications, it will be all the more important to have effective and appropriately flexible accountability mechanisms built into them.

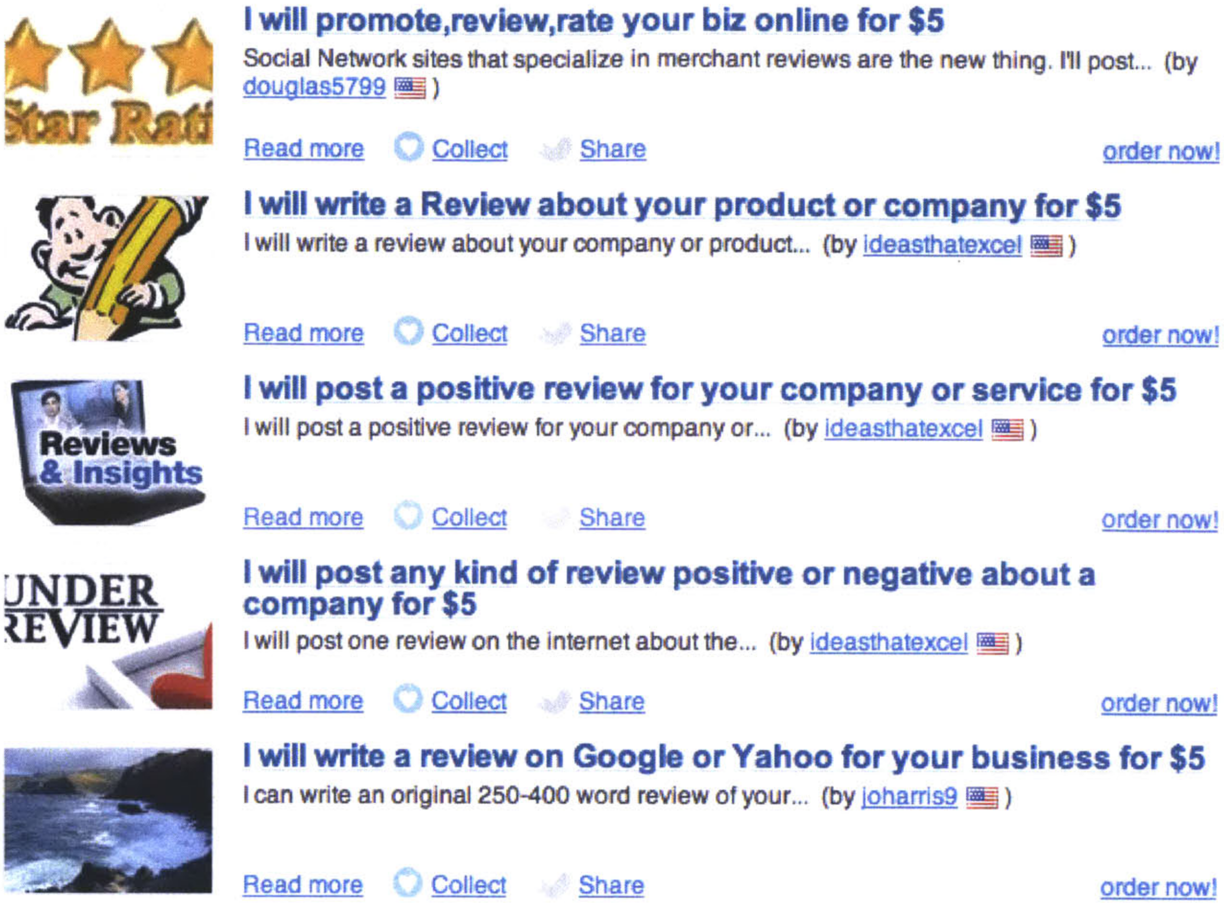
Chapter 6

Yelp

> It's a 75 dollar flat fee to sit down in this place. I ordered the Pork Ravioli. First off, is this a restaurant for anorexics? because the portion sizes are for hobbits. How did it taste? BLEH! ... heed my words and do NOT step into this greedy deception they call a restaurant.
>
> —Review of restaurant Anella on Yelp.com by user James D.

When Blair Papagni saw this online review of her New York restaurant, she immediately contacted popular review site Yelp.com and asked them to remove it, explaining that the restaurant she owns neither charges a $75 fee to seat customers nor offers pork ravioli on its menu. Yelp ultimately denied Papagni's request, responding that they would leave the review up because it "appears to reflect the personal experience and opinions of the reviewer." Despite this refusal, identifying and filtering out falsified reviews has become increasingly crucial to review aggregation sites like Yelp and Amazon as the influence of their ratings on consumer behavior grows steadily and companies consequently develop more schemes for inflating their own online reputations or besmirching those of their competitors with fake reviews (Kludt, 2011). Like e-mail, online applications that collect user reviews can suffer from traditional commercial spam postings, also referred to as "disruptive opinion spam." This type of disruptive spam, consisting primarily of advertisements, questions, and other irrelevant or non-opinion text, can be irritating to users but it is relatively easy for sites like Yelp to detect and delete (Ott, Choi, Cardie, & Hancock, 2011). A more difficult problem for these sites is how to deal with "deceptive opinion spam," or false reviews that have been deliberately authored to seem authentic. Yelp, like Wikipedia, is concerned with trying to verify the accuracy of the massive volume of information provided by its users, but unlike Wikipedia, there is no objectively correct information that user entries can be checked against. Identifying a fake restaurant review on Yelp, in other words, is a considerably harder and more subjective proposition than detecting factually inaccurate Wikipedia entries.

On its surface, Papagni's outrage over a few negative jabs about her restaurants portion sizes and prices may seem disproportionate to the seriousness of the offense, but while fake online reviews may at first seem like a trivial problem they are a signif-



I will promote,review,rate your biz online for $5

Social Network sites that specialize in merchant reviews are the new thing. I'll post... (by douglas5799)

I will write a Review about your product or company for $5

I will write a review about your company or product... (by ideasthatexcel)

I will post a positive review for your company or service for $5

I will post a positive review for your company or... (by ideasthatexcel)

I will post any kind of review positive or negative about a company for $5

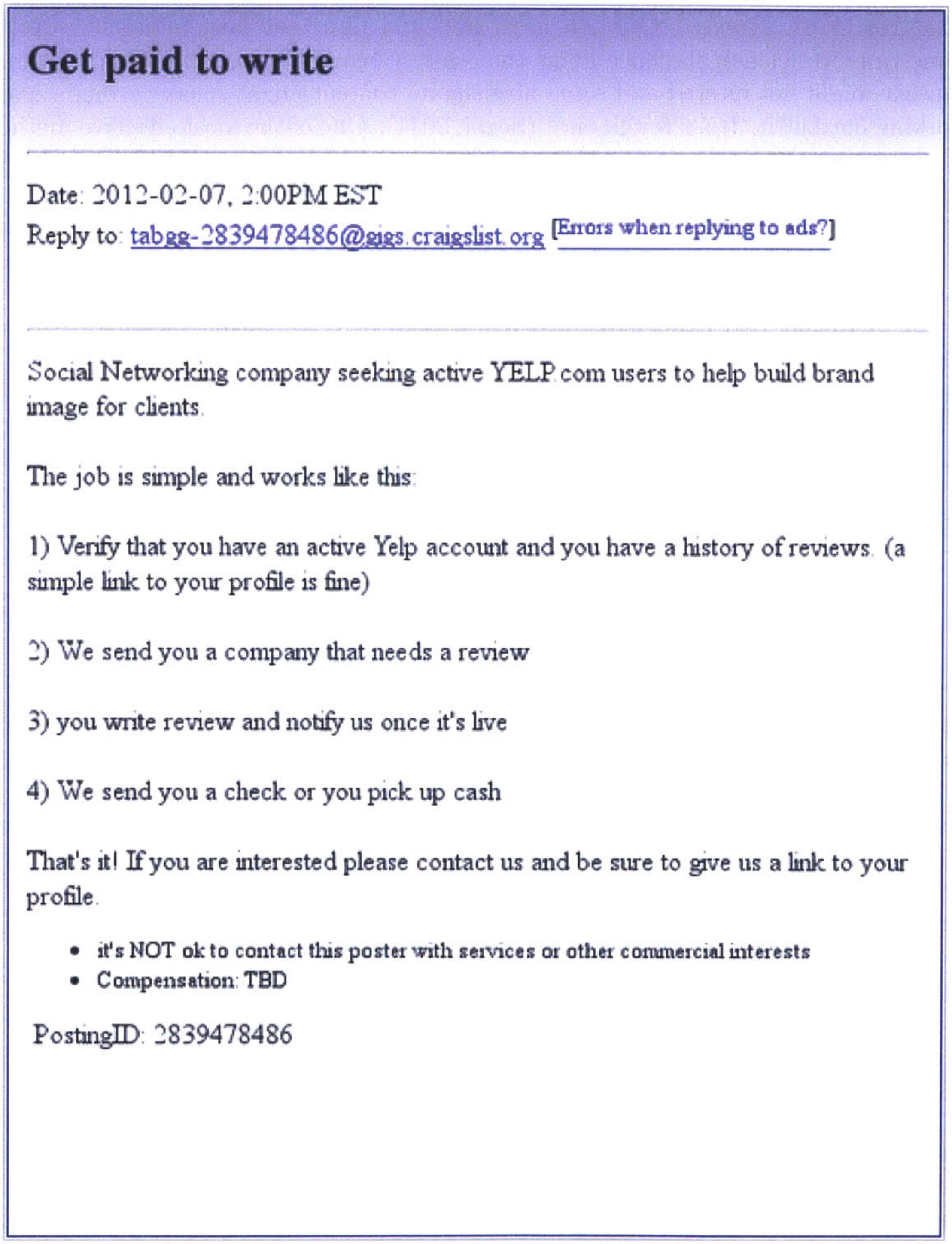
I will post one review on the internet about the... (by ideasthatexcel)

I will write a review on Google or Yahoo for your business for $5

I can write an original 250-400 word review of your... (by joharris9)

Figure 6-1: Websites like Fiverr feature numerous users willing to post fake reviews in exchange for payment.

icant and growing concern for businesses and review sites alike. From the perspective of the companies whose products and services are reviewed online, the ratings and comments provided on the most popular review sites can have a major impact on business. One recent study of restaurants in Washington found that a one-star increase in Yelp rating led to between a 5 and 9 percent increase in revenue (Luca, 2011), thus businesses have clear economic incentives for trying to boost their own online ratings above those of their competitors. This happens most often in the form of companies paying people either to write glowing reviews of their business and products or to write damningly negative reviews of their competitors, or both. As the influence of online review sites increases, so too does the demand for fake reviews, to the point where micro-employment services like Mechanical Turk, Craigslist, and Fiverr are flooded with advertisements by people offering to write them, as shown in Figure 6-1, and University of Illinois computer science professor Bing Liu estimates that as many as 30 percent of online reviews may be fake for especially competitive services and products (Weise, 2011). One February 2012 post on Craigslist, shown in Figure 6-2, explicitly solicits Yelp users to post fake reviews, requiring that all interested applications verify that they have active Yelp accounts and a history of



Get paid to write

Date: 2012-02-07, 2:00PM EST

Reply to: tabgg-2839478486@gigs.craigslist.org [Errors when replying to ads?]

Social Networking company seeking active YELP.com users to help build brand image for clients.

The job is simple and works like this:

1) Verify that you have an active Yelp account and you have a history of reviews. (a simple link to your profile is fine)

2) We send you a company that needs a review

3) you write review and notify us once it's live

4) We send you a check or you pick up cash

That's it! If you are interested please contact us and be sure to give us a link to your profile.

- it's NOT ok to contact this poster with services or other commercial interests
- Compensation: TBD

PostingID: 2839478486

Figure 6-2: A February 2012 Craiglist posting offering to pay Yelp users for posting false reviews.

reviews on the site.

With the credibility and usefulness of their reviews at stake, sites like Yelp have taken steps to implement filtering algorithms and remove the false reviews they flag but these efforts have met with mixed success. Yelp spokesman Vince Sollitto told *The New York Times*: "Our job is to find and filter out fake reviews ... At the same time we let our audience know that this system isn't perfect. Some legitimate content might get filtered and some illegitimate content might sneak through. We're working hard at it. It's a tough one" (Segal, 2011). One of the most effective methods of mitigating these fake reviews has nothing to do with Yelp's sophisticated filtering algorithms, however. Hinted at in the specifications required by the social networking company on Craigslist, this protection lies in the way Yelp users build up their online identities and establish reputations within the site. Yelp's reputational mechanisms are not fool-proof; it is still possible to solicit established Yelpers to write fake reviews (as the Craigslist ad explicitly does) but it does pose significant obstacles to those wishing to game the system and, perhaps even more importantly, it allows users to individually assess the authors of reviews they read and decide how credible and relevant those authors' opinions are to their own preferences. This chapter will focus on Yelp's reputation system, examining how it is implemented, the ways in which it benefits Yelp as well as Yelp users, possible threats to its integrity, and its overall impacts on account discardability and user accountability.

6.1 The Role of Online Reputation Systems: Sanctioning and Signaling

Online reputation systems were originally developed not for review sites but for commercial Internet applications, where users were purchasing items from strangers and worried about how they should decide whether to trust these unknown sellers. In particular, buyers were concerned that they might pay for a product or service but never actually receive it from the seller and have no recourse for holding that seller accountable without knowing his identity. Schneier (2006) points out that "In an anonymous commerce system – where the buyer does not know who the seller is and vice versa—it's easy for one to cheat the other. This cheating, even if only a minority engaged in it, would quickly erode confidence in the marketplace."

6.1.1 eBay & The Origins of Online Reputation

In response to these concerns about unknown buyers and sellers cheating each other, auction website eBay instituted a reputation system that allowed users to provide feedback about their transactions with other users. After two users completed an auction, the buyer and the seller could each rate the other as either a +1 (positive), 0 (neutral), or a -1 (negative) and also leave a brief comment about their experience. Users were only allowed to enter these rankings after successfully completed auctions, and eBay then calculated reputation scores based on how many positive rankings a user had gotten, minus the negative rankings they had received. This score was then automatically displayed on each user's auction page, along with individual comments and a more detailed breakdown of the rankings over time. New users entered the site with a reputation score of zero, and were denoted by a small sunglasses icon displayed next to their username during the first month of their membership; thus, users who decided to discard their old identities due to negative reputations were forced to start over entirely and build a new reputation from scratch (Resnick, Zeckhauser, Swanson, & Lockwood, 2003). This reputation system allowed sellers to retain their anonymity online while still enabling them to persuade buyers that they could be trusted. Schneier (2006) notes, "eBay's feedback system doesn't work because there's a traceable identity behind that anonymous nickname. eBay's feedback system works because each anonymous nickname comes with a record of previous transactions attached, and if someone cheats someone else then everybody knows it."

Studies of eBay's system demonstrated clearly that the seller ratings impacted how successful their business was on eBay as well as what prices they could charge. Resnick et al. (2003) estimated that established sellers with strong positive reputations could charge prices that were roughly 8.1 percent higher prices than new sellers offering the same merchandise. The researchers concluded that "eBay's public reputation scores play a significant role in the marketplace, and ... virtually all significant sellers have strong reputations. In our controlled experiment, a seller with a strong reputation received a price premium, even holding constant quality of goods, skill at listing, and responsiveness to inquiries." In other words, the online reputations mattered; the positive rankings from other users were worth something—in financial terms—to the eBay sellers. More importantly, eBay's model demonstrated that it was possible to effectively recreate in an online context two crucial elements of reputation that we rely on in the real world: a history of past interactions and an expectation that that history will influence our future interactions.

6.1.2 The Shadow of the Future

When we develop real-world relationships with other people, trust usually develops over time, based on two key features. First, we can decide how much we trust other people based on our shared history of past interactions. Second, we are encouraged to trust others by our conviction that how they behave now will affect our future interactions—good behavior will likely be reciprocated at some point while there may be retaliation for any malicious behavior. Political scientist Robert Axelrod terms this phenomenon the "shadow of the future" and argues that this shadow effectively constrains people's present behavior because of the expectation that this behavior will influence their future interactions with others (Resnick, Kuwabara, Zeckhauser, & Friedman, 2000). The eBay reputation system provided users with both of these features: a history of the past interactions of complete strangers and a confidence that sellers would be unlikely to cheat or defraud them for fear that such an action would negatively affect their ranking and thereby influence their ability to sell more items in the future. Resnick et al. (2003, p. 80) explain: "Because people know that their behavior today will affect their ability to transact in the future, not only with their current partner but with unknown others as well, opportunistic behavior is deterred. Moreover, less reliable players are discouraged from joining the marketplace." In other words, reputation schemes can allow users to make reasonable, instantaneous assessments about how much they could trust complete strangers.

In some respects, online applications like eBay and Yelp have even been able to improve on standard notions of reputation. Although the anonymous identities and lack of face-to-face interaction are obvious drawbacks of establishing reputations in cyberspace, Internet markets also have some advantages in this arena. It is relatively cheap and easy for websites to continuously collect and aggregate new data about their users, including comments and rankings from other users as well as objective statistics about how long someone has been a member, how many interactions they've completed, and more. Additionally, it is equally cheap and easy to share that information with all of the site's viewers. By contrast, assembling that much information for real-world businesses and transmitting it to millions of individuals would be a much more expensive and time-intensive process. "The Internet can vastly accelerate and add structure to the process of capturing and distributing information," Resnick et al. (2000, p. 47) write of online reputation systems, adding that "the same technology facilitating market-style interaction among strangers also facilitates the sharing of reputations that maintain trust."

6.1.3 Moral Hazard & Adverse Selection

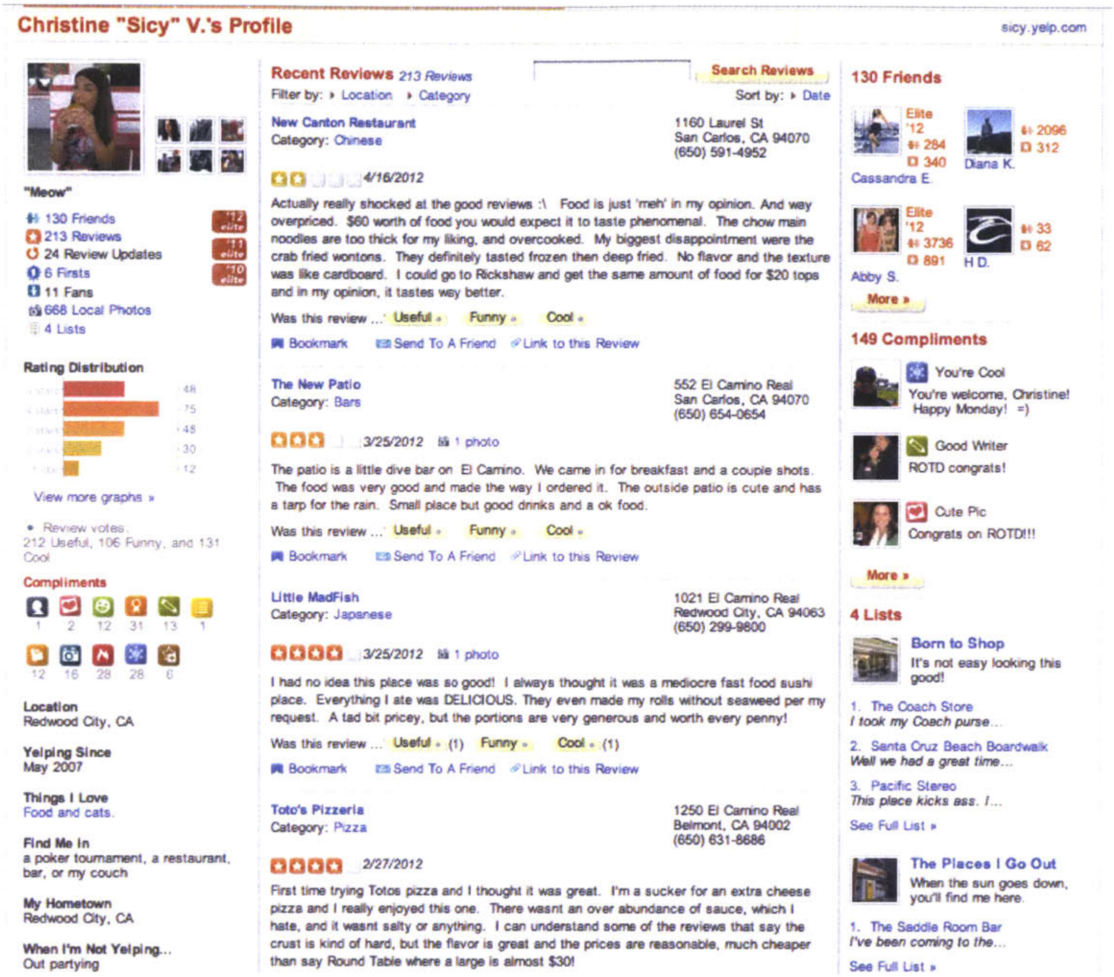
Though all reputation systems aim to help users assess each other in the context of different online communities, these reputational mechanisms come in many varied forms and can serve several different purposes, depending on the functions and characteristics of the Internet applications they are designed for. Two overarching problems they can be used to address in online applications are moral hazard and adverse selection. Moral hazard occurs when actors take undue risks because they do not bear the full negative consequences of their actions. For instance, people with health insurance might not take measures to avoid injury knowing that they will not bear the costs of their medical care. Similarly, in an online setting, a seller on eBay might decide not to send a purchased product to a buyer after receiving payment for it, since, in the absence of any reputation system, the seller would bear no cost for this fraudulent behavior. As illustrated by the eBay example, online reputation systems can mitigate the problem of moral hazard by sanctioning actors for their misbehavior, effectively imposing costs on their actions that would otherwise have been borne by others. Thus, if eBay sellers expect that cheating their customers will diminish their future sale volume by damaging their online reputation, thereby losing them more money in the long-term than they gain from it in short-term, they are less likely to engage in morally hazardous behavior.

Similarly, many online applications are at risk for adverse selection, in which users have access to asymmetric information so they cannot differentiate between good and bad products or services being offered, and therefore the bad products (which can be offered at lower prices) end up driving the good ones out of the market. In e-commerce, adverse selection issues arise when buyers are unable to assess the quality of multiple competing products (e.g. electronics) or opinions (i.e., reviews) being offered on a site. Reputation systems can help solve the information asymmetries that cause adverse selection by providing signals to users about the quality of the products or opinions they are trying to assess. These signals can take a variety of different forms, ranging from online reviews of digital cameras that help customers assess the quality of those cameras to information about the authors of those reviews (e.g. how many other reviews they've written) that help other users assess the reliability and quality of the opinions they are reading. In other words, publishing user reviews and publishing statistics about the reviewers are two different ways of trying to address the information asymmetries that confront online users by using reputation systems as signaling mechanisms.

Yelp addresses both moral hazard and adverse selection concerns by publishing user-written reviews about many different businesses along with a great deal of reputational information about its users, ranging from how long they've been Yelp members, to how many reviews they have written, to the frequency with which they assign each star rating, to the number of other users who have judged their reviews to be helpful. This information is contained in detailed user profiles, as shown in Figure 6-3. The reviews themselves help prevent moral hazard at businesses, like restaurants, where customers who receive poor service or subpar meals will still be expected to pay full price at the end of the meal. However, the risk that restaurants might be subject to moral hazard and cut corners knowing their customers would still have to pay is mitigated by the restaurants' fear that dissatisfied customers will write negative reviews and adversely affect their future business. Additionally, Yelp reviews help users identify good and bad restaurants, alleviating the information asymmetries about restaurant quality that might cause adverse selection. Thus, Yelp reviews serve as both a potential sanction to proprietors in the event of poor service and also as signals to customers about which businesses they should and should not frequent.

In addition to these functions of helping users identify good businesses and enjoy good service at them, Yelp's reputation system addresses the moral hazard and adverse selection problems inherent in the reviewing process itself. For instance, moral hazard could be a problem for reviewers who believe there will be no costs to themselves for authoring falsely positive or negative reviews; however, their online reputations could suffer from giving too many one- or five-star reviews (a metric that is often cause for reviews being filtered out of Yelp) or from other users labeling their reviews as inaccurate or unhelpful. Similarly, information asymmetries could prevent readers from distinguishing the quality and reliability of any given user's reviews, but the reputational data helps mitigate the risks of adverse selection regarding which reviews to trust. Thus, the reputational data published about Yelp users can serve as both a sanction for misbehaving reviewers and also a signal to other users about that reviewer's reliability.

Yelp's reputation system is not just valuable as a sanctioning and signaling mechanism, however, it also serves another, distinct function in helping users find businesses and reviewers that will most closely match their own personal preferences. After all, not everyone likes the same restaurants and part of the value of Yelp is not just the ability to see overall rankings of businesses but to search for specific features, from cuisine style to price category, that suit the desires of a given user. Similarly, users



Christine "Sicy" V.'s Profile

sicy.yelp.com

"Meow"

130 Friends
213 Reviews
24 Review Updates
6 Firsts
11 Fans
668 Local Photos
4 Lists

'12 elite
'11 elite
'10 elite

Rating Distribution

View more graphs »

- Review votes.
212 Useful, 106 Funny, and 131 Cool

Compliments

Location
Redwood City, CA

Yelping Since
May 2007

Things I Love
Food and cats.

Find Me In
a poker tournament, a restaurant, bar, or my couch

My Hometown
Redwood City, CA

When I'm Not Yelping...
Out partying

Recent Reviews *213 Reviews*

Search Reviews

Filter by: ▸ Location ▸ Category

Sort by: ▸ Date

New Canton Restaurant
Category: Chinese

1160 Laurel St
San Carlos, CA 94070
(650) 591-4952

4/16/2012

Actually really shocked at the good reviews :\ Food is just 'meh' in my opinion. And way overpriced. $60 worth of food you would expect it to taste phenomenal. The chow main noodles are too thick for my liking, and overcooked. My biggest disappointment were the crab fried wontons. They definitely tasted frozen then deep fried. No flavor and the texture was like cardboard. I could go to Rickshaw and get the same amount of food for $20 tops and in my opinion, it tastes way better.

Was this review ...' Useful Funny Cool

Bookmark Send To A Friend Link to this Review

The New Patio
Category: Bars

552 El Camino Real
San Carlos, CA 94070
(650) 654-0654

3/25/2012 1 photo

The patio is a little dive bar on El Camino. We came in for breakfast and a couple shots. The food was very good and made the way I ordered it. The outside patio is cute and has a tarp for the rain. Small place but good drinks and a ok food.

Was this review ...' Useful Funny Cool

Bookmark Send To A Friend Link to this Review

Little MadFish
Category: Japanese

1021 El Camino Real
Redwood City, CA 94063
(650) 299-9800

3/25/2012 1 photo

I had no idea this place was so good! I always thought it was a mediocre fast food sushi place. Everything I ate was DELICIOUS. They even made my rolls without seaweed per my request. A tad bit pricey, but the portions are very generous and worth every penny!

Was this review ... Useful (1) Funny Cool (1)

Bookmark Send To A Friend Link to this Review

Toto's Pizzeria
Category: Pizza

1250 El Camino Real
Belmont, CA 94002
(650) 631-8686

2/27/2012

First time trying Totos pizza and I thought it was great. I'm a sucker for an extra cheese pizza and I really enjoyed this one. There wasnt an over abundance of sauce, which I hate, and it wasnt salty or anything. I can understand some of the reviews that say the crust is kind of hard, but the flavor is great and the prices are reasonable, much cheaper than say Round Table where a large is almost $30!

130 Friends

Elite '12
284
340
Cassandra E.

2096
312
Diana K.

Elite '12
3736
891
Abby S.

33
62
H D.

More »

149 Compliments

You're Cool
You're welcome, Christine! Happy Monday! =)

Good Writer
ROTD congrats!

Cute Pic
Congrats on ROTD!!!

More »

4 Lists

Born to Shop
It's not easy looking this good!

1. The Coach Store
I took my Coach purse...

2. Santa Cruz Beach Boardwalk
Well we had a great time...

3. Pacific Stereo
This place kicks ass. I...

See Full List »

The Places I Go Out
When the sun goes down, you'll find me here.

1. The Saddle Room Bar
I've been coming to the...

See Full List »

Figure 6-3: A Yelp user's profile contains extensive reputation information such as how long the user has been a Yelp member, how many reviews she has written, the rating distribution of those reviews, and the number of compliments those reviews have received from other users.

may decide which reviews to trust based on how much they feel they have in common with certain reviewers, either based on those reviewers' ratings of other restaurants, their review-writing styles, their self descriptions, or the characteristics of businesses that they single out for discussion in their reviews. In other words, Yelp's reputational profiles can serve as a sanctions, signals, and customization tools to help individual users match their preferences to those of specific other reviewers.

6.2 Benefits of Reputation Systems

There are many advantages to reputation systems from the perspective of end-users, who can use them to more easily and reliably assess the quality of products or opinions as well as enjoy online prestige of their own by building up strong reputations. Such mechanisms are also increasingly valuable to application owners and operators looking to build trust between users, filter out lower-quality information, match users who vary in their interests and tastes, and establish user loyalty to their applications (Dellarocas, 2010).

Based on an application designer's priorities, a reputation system can emphasize different combinations of these four functions—trust-building, filtering, matching, and user lock-in—depending on what reputational data is collected and how it is displayed to other users. For instance, establishing trust between buyers and sellers was the primary goal of eBay's reputation system. This motivated the simple numerical scoring system for transactions, based on the underlying assumption that the qualities of a reliable seller were generally objective and shared by all buyers and therefore did not require extensive detailed comments for more personalized matching of buyers and sellers (the comments shared by users about each other were limited to one line in length). Review aggregation sites like Yelp, however, emphasize the filtering and matching functions in the design of their reputation systems by encouraging users to evaluate the quality of other users' reviews and soliciting longer, more detailed textual input from reviewers while de-emphasizing the importance of individual numerical rankings. These strategies align well with their main priorities of being able to distinguish between higher and lower quality reviews and helping users find products and services that most closely match their own tastes and preferences. By contrast, reputation systems for gaming applications like World of Warcraft and Xbox Live are designed with a greater focus on matching players together and establishing user loyalty and retention in the highly competitive environment of online games.

User lock-in can be a particularly strong incentive for designers to include reputation mechanisms in their applications. Since reputations are usually tied to specific application, users who have built up established reputations within an application like Yelp are less likely to defect to a competing site where they would have to start all over again building up their reputations. Besides reducing user attrition, individuals' reluctance to constantly create new reputations from scratch can also be a powerful means of making their online identities less discardable. However, while users with positive reputations may take care not to misbehave in any ways which could jeopardize their established credentials, those who have negative reputations have nothing to lose by jettisoning their current identities and starting over with fresh ones. This is just one of several user activities that can undermine the effectiveness and utility of anonymous reputation systems.

6.3 Risks of Reputation Systems: Whitewashing, Sybils, and Lying

To operate effectively, a reputation system must exhibit at least three properties, not all of which are easily replicated in cyberspace. First, such a system must have long-lived entities or identities that carry an expectation of future interaction with others in order to inspire the "shadow of the future" effect. Second, the system must collect and distribute feedback about current interactions, and finally, that feedback must be used to inform future interactions and guide trust decisions between users (Resnick et al., 2000). As discussed earlier, Internet applications have some advantages over the reputation systems in the physical world when it comes to the latter two properties—it is much simpler and more reliable to aggregate and display large volumes of user feedback to a broad audience online than it is by word of mouth. However, the discardability of Internet identities makes the first criteria much more challenging to meet for application designers and can pose a serious threat to the integrity of online reputation systems.

Applications like eBay and Yelp run the risk that users who rack up extremely negative reputations—either by cheating customers in online auctions or writing falsified reviews—will simply create new accounts and shed their old negative reputational data, a process known as "whitewashing" (Friedman, Resnick, & Sami, 2007). As in other Internet applications, if it is completely costless to create a new identity—and thereby a new reputation—on Yelp, the reputational information provided becomes much less meaningful and the users much less accountable. Another, related threat to the effectiveness of online reputation systems is the possibility of an individual user creating hundreds of "sybils," or ghost identities, for the sole purpose of boosting the reputation of their primary identity. On Yelp, this phantom feedback problem could manifest either as a reviewer creating multiple accounts to boost the usefulness ratings of their own reviews or as a business owner creating accounts to lavish glow reviews on their own business (Friedman et al., 2007). Phantom feedback overlaps to a certain extent with a third serious threat to reputation systems: false feedback, in which users provide incorrect feedback either because they are being paid by a third party to propagate a specific opinion or because they have other incentives to lie. Whitewashing reputations, phantom feedback, and false feedback can greatly undermine the effectiveness and utility of online reputation systems; however, application designers can (and do) mitigate the consequences of all three of these problems by carefully crafting reputation systems that are specifically suited to the purposes of their application and resilient to anticipated forms of user manipulation.

To combat whitewashing, reputation systems can either require unique identities by individuals, so that they cannot create multiple accounts, or impose some cost—financial or otherwise—on the creation of new identities. Friedman et al. (2007) have proposed a system of unreplaceable pseudonyms, or "once-in-a-lifetime identifiers," intended to prevent individual users from creating multiple accounts within a reputation system. They suggest issuing an anonymous certificate to each user containing a single identifier that is unrelated to the user's real identity but can be used to prevent that person from creating more than one account in the same reputation system. Such a system would greatly diminish the risks of whitewashing and phantom feedback, but would also require the establishment of a trusted third party for issuing certificates and potentially pose some concerns about the anonymity of the proposed once-in-a-lifetime identifiers.

In the absence of certificate-based unreplaceable pseudonyms, which most online reputation systems have not adopted thus far, the only remaining means of tackling whitewashing and phantom feedback is to impose some cost on users creating new accounts. For instance, new users can be forced to "pay their dues" for a certain period of time, until their reputations are deemed sufficiently established to be considered trustworthy. "Game-theory analysis demonstrates that there are inherent limitations to the effectiveness of reputation systems when participants are allowed to start over with new names," write Resnick et al. (2000, p. 48). "In particular, newcomers (those with no feedback) should always be distrusted until they have somehow paid their dues, either through an entry fee or by accepting more risk or worse prices while developing their reputations," they conclude. Review site Angie's List, which serves a similar function as Yelp in aggregating user reviews, takes the approach of charging its users an annual subscription fee of roughly $60 to read and write reviews on its site. Though its business model stands in stark contrast to Yelp's, which does not charge users any fees, Angie's List has also enjoyed considerable success, with more than one million paying users (Lieber, 2012). Unlike Yelp, which focuses primarily on reviews of bars and restaurants, Angie's List specializes in reviews of "high-cost-of-failure" businesses, such as home remodeling and medical services, where a reliable recommendation is most critically important to consumers. In these cases, users may be reassured by the knowledge that, although they cannot view the names of the reviewers on Angie's List, the website has a record of every user's real name and credit card information. The site itself lauded this lack of total anonymity in its 2011 securities filing, writing, "The anonymity of the Internet renders it inherently susceptible to outright manipulation by unscrupulous service providers and unhappy customers, so consumers have limited means for discerning which information they should trust" (Lieber, 2012).

Yelp users, who have less to fear from reading an inaccurate review of a restaurant, tolerate a fairly high degree of anonymity, however, instead relying on the site's reputational metrics to discern which reviewers they can trust. The site itself is actively engaged in trying to regulate user behavior using this same reputation information. Indeed, although the specific criteria for Yelp's filtering algorithm are kept secret by the company to prevent users from manipulating their reviews to evade the filter, Yelp has stated that the primary focus of the algorithm is how well established a reviewer's reputation is. Lowe (2010), Yelp's Director of Outreach & Public Policy, explained in a posting on the company's blog that the site's filtering algorithm "syndicates established users' reviews from their user pages onto business pages. This automated process sometimes creates the perception that reviews are being deleted and re-added over time; what's actually happening is users are becoming more-or-less established over time." In other words, Yelp uses some of the same elements of its reputation system that prevent whitewashing and sybil accounts to also identify and filter out false feedback, by targeting users with less well established reputations. Notably, in spite of its paying users, false feedback is also a concern for Angie's List, which employs a team devoted to investigating "problematic reviews" through algorithms and human intervention (Lieber, 2012).

Researchers have identified other methods of detecting fake online reviews, besides a reviewer's reputation. One 2011 study was able to develop an automated classifier that could correctly identify nearly 90 percent of fabricated hotel reviews based solely on linguistic cues. For instance, Ott et al. (2011) noted, "We find increased first person singular to be among the largest indicators of deception, which we speculate is due to our deceivers attempting to enhance the credibility of their reviews by emphasizing their own presence in the review." Ott told *Bloomberg Businessweek* that truthful reviews were much more likely to describe the physical space of the hotel using specific nouns and adjectives. "Since spammers weren't familiar with the look of the hotel, they spent more time talking about themselves, the reasons they went on a trip, and their traveling companions," Ott explained, adding that "the word 'husband' is very indicative of deception" (Weise, 2011).

Combining linguistic cues with reputational data is essential for identifying false feedback in reputation systems because, as reputation systems become more robust, retailers are developing sneakier ways to game them. Where once a chef might just have created dozens of accounts himself to write five-star reviews of his restaurant, business owners are beginning to understand the importance of soliciting these positive reviews from customers or other people, with more established online reputations, and have begun offering up to $80 for reviews from trusted Yelp users who had achieved the site's "Elite status," it's highest reputational distinction (Weise, 2011). Some retailers have even experimented with offering their customers price rebates in exchange for their posting online reviews. An English hotel, The Cove, reportedly gave its customers a 10 percent discount for posting "honest but positive" reviews of the hotel on popular travel site TripAdvisor (Streitfeld, 2011). More recently, the company VIP Deals drew the attention of the U.S. Federal Trade Commission (FTC) for offering customers who purchased its Kindle e-reader cases a full refund in exchange for writing a review of its product on Amazon. In the letter offering customers this refund, the company hinted strongly at what type of reviews it was looking for, writing: "We strive to earn 100 percent perfect 'FIVE-STAR' scores from you!" Notably, the scheme was successful, with *The New York Times* reporting that, by the time the rebate offer ended in late January 2012, the VIP Deals Kindle case was "receiving the sort of acclaim once reserved for the likes of Kim Jong-il. Hundreds of reviewers proclaimed the case a marvel, a delight, exactly what they needed to achieve bliss. And definitely worth five stars" (Streitfeld, 2012).

FTC associate director for advertising practices Mary Engle said the FTC was "very concerned" about the ethics of offering incentives for positive reviews, noting that, "Advertising disguised as editorial is an old problem, but it's now presenting itself in different ways." Still, while federal regulators may be able to impose some measures to discourage such behavior, given the prevalence of the problem and the international nature of Internet application users, the primary burden of identifying and sanctioning false feedback is likely to continue to fall on the application operators who design and maintain reputation systems. Amazon has its own program to encourage more high-quality product reviews on its site by inviting selected users to join its Amazon Vine program and become "Vine Voices" (the rough equivalent of Yelp's elite status). Vine Voices are "selected based on the trust they have earned in the Amazon community for writing accurate and insightful reviews" as indicated primarily by the number of helpful votes they receive from other users, according to Amazon. These selected reviewers are given free advance products to review and their reviews, like those posted by Yelp elite users, are designated with a special symbol indicating the status of the reviewer. Though retailers do pay Amazon for the privilege of being able to release advance products to Vine Voices, the primary function of the program is to generate high-quality reviews for the site, not to make money, Amazon vice president Jeff Belle asserts (Springen, 2009).

6.4 Persistent Online Persona

Yelp's reputation system serves multiple functions within the application; it is a means of sanctioning misbehaving users for false or unhelpful reviews, a means of signaling to other users the reliability of reviewers as well as how closely reviewers' individual tastes and preferences match those of the readers, a means of rendering user identities less easily discardable by forcing reviewers to invest time and energy in building up their reputations, and finally, a means of filtering out potentially fake or unreliable reviews from less well established users. The combination of these effects can be a powerful tool for promoting accountability online in a purely pseudonymous system, without tying Internet identities back to their real-world counterparts. In other words, reputation systems are an important component of several applications within the "anonymous-and-accountable" quadrant of the anonymity-accountability axes since, as (Donath, 1999, p. 54) points out, "in an electronic environment in which pseudonyms are prevalent, only the sanctions that do not require a connection to the real world are practical."

Still, some researchers have raised concerns about how anonymous such reputational mechanisms can be, when they are based on the ongoing collection of data about an individual. Bethencourt, Shi, and Song (2010) have argued that "in all such systems, a user is linked by their pseudonym to a history of their messages or other activities . . . recent work has shown that very little prior information about an individual is necessary to match them to their pseudonym. Building a truly private forum requires abandoning the notion of persistent identities." To protect anonymity more strongly within reputation systems, this research team proposed a system of encrypted "signatures of reputation" that could be used to sign users' opinions or feedback but, instead of verifying their identities like a standard digital signature, these reputation signatures would instead only verify the reputational data associated with the author. Such a scheme, they suggest, could allow application designers and users to derive many of the benefits of existing reputation systems while providing much stronger anonymity protections to end users. Bethencourt et al. (2010) write:

> We might imagine an anonymous message board in which every post stands alone—not even associated with a pseudonym. Users would rate posts based on whether they are helpful or accurate, collect reputation from other users' ratings, and annotate or sign new posts with the collected reputation. Other users could then judge new posts based on the author's reputation while remaining unable to identify the earlier posts from which it was derived. Such a forum would allow effective filtering of

> spam and highlighting of quality information while providing an unprecedented level of user privacy.

However, it is not clear that such a signature-based reputation system would enable users or application designers to enjoy any of the matching functionalities of identity-based reputation systems, like Yelp's, that help users assess how well suited feedback would be to their personal preferences. Eliminating the user identities associated with online reputations and replacing them solely with aggregated reputation data might also diminish users' ability to build ties to other members of online communities and develop trust relationships with them. As Goffman (1959, p. 75) observes, belonging to a society requires stability of self-presentation and "standard-maintaining routines." He notes that "a status, a position, a social place is not a material thing, to be possessed and then displayed; it is a pattern of appropriate conduct, coherent, embellished, and well articulated." The power of online reputation systems lies in their ability to establish these distinct and persistent personae in cyberspace, where identities are often too transient and discardable to be meaningful. In doing so, reputation systems can bolster many elements of Internet application communities, by building trust between individual end-users and ensuring that they can be held accountable for their online actions, even when they act anonymously.

Chapter 7

Design Patterns for Identity Investment-Privilege Trade-offs

> Whether one is for or against anonymity online, a design decision bearing on it, made first as an engineering matter, can end up with major implications for social interaction and regulation.
>
> —Zittrain (2009, p. 33)

The case studies explored in the previous chapters highlight several observations about the nature of Internet applications, including the rich diversity of forms they take and functions they serve, as well as the equally rich diversity of ways in which their intended functions can be abused or distorted by users to engage in malicious behaviors ranging from harassment to fraud. The identity mechanisms employed by these different applications reveal the complicated interplay between anonymity and accountability in virtual identities and the variety of actors positioned at different control points in each of these applications. Application designers, in particular, have tremendous power to dictate what kinds of identities their users can create, what degree of accountability and anonymity will be embedded in these identities, and the ways in which other actors are—or are not—empowered to help hold their users accountable. Yelp's extensive reputational profiles, Second Life's hierarchy of paid and unpaid avatars, e-mail's lack of user authentication, Mathew Firsht's ability to retrieve from Facebook the IP addresses of every computer that accessed his false profile page—all of these features stem directly from the application's architecture and the decisions made by its designer. Implementing accountability at the application layer means that application designers control, to a great extent, the means and mechanisms by which Internet identities can and cannot be held accountable.

7.1 The Role of Application Design Patterns

The implications of this application-layer approach to accountability can be both positive and negative. The primary benefit of this approach is that it enables designers to tailor-make suitable accountability mechanisms for the specific functions and technical architecture of their respective applications, rather than being subject to a single, one-size-fits-all network-layer approach to accountability. The sheer number of new and existing Internet applications highlights the importance of allowing for this diversity of accountability schemes, but it also suggests the potential negative ramifications of such a system of customized, application-specific mechanisms: the responsibility for implementing accountable identity schemes rests largely on the application designer, who may or may not pay attention to accountability and make appropriate design decisions regarding its implementation. In other words, the advantage of such an approach is that individual application designers are empowered to design accountable identity mechanisms but the disadvantage is that some of them likely won't, either due to lack of interest or lack of knowledge, thereby leaving their applications vulnerable to rampant user misbehavior.

To help mitigate the drawbacks of relying on application designers to implement accountability themselves, the final chapters of this thesis present several design patterns specifically aimed at promoting user accountability in Internet applications. Describing the ongoing "tussles" in the Internet space between different stakeholders with opposing interests, Clark, Wroclawski, Sollins, and Braden (2005, p. 472) encourage this design pattern approach, writing: "If application designers want to preserve choice and end user empowerment, they should be given advice about how to design applications to achieve this goal ... we should generate 'application design guidelines' that would help designers avoid pitfalls, and deal with the tussles of success." The design guidelines laid out in this and subsequent chapters are drawn primarily from the earlier case studies of accountability mechanisms implemented in existing applications. They are intended to ease the burden that application-layer accountability places on application designers, by providing some background and initial guidance on different ways to embed accountability and anonymity in online identities, depending on the function, structure, and target user base of an application.

7.2 The Investment-Privilege Trade-Off

Suppose two Yelp users each post a review for the same new restaurant. One of the Yelpers has been a member of the site for nearly two years and has contributed more than 100 reviews of different businesses, many of which have been rated "useful" and "cool" by other users, and she recently achieved "elite" status on the site; her review is the first one listed on the restaurant's Yelp page. The other user created her Yelp account one month ago and has only posted two reviews, neither of which have received any complimentary ratings; her review of the same restaurant is automatically filtered by Yelp's algorithms and hidden from view on the restaurant's main page.

Now, consider two Second Life avatars, one of whom owns a popular night club and maintains a careful list of other users who are banned from his club due to past transgressions, only allowing entry to avatars who have paid or verified accounts. He earns upwards of 10,000 Linden dollars (or roughly $40) per month by charging patrons for drinks and cover fees, easily turning a profit even after paying the monthly $9.95 account fee to Linden Lab. Another avatar is controlled by a free account holder and occasionally wanders through the virtual world, exploring the different areas open to the public and carrying on conversations with some of the other users he encounters, but never purchasing any Linden dollars or participating in the Second Life economy.

Finally, recall the certified e-mail system that Yahoo and AOL experimented with in 2006 when they announced that they would allow some bulk e-mails to bypass their spam filters, so long as the sender (or sending company) was willing to pay them a fraction of a penny per message. Under this system, larger, more successful organizations with sizable marketing budgets—including the White House and Target—were able to pay to ensure that their messages would arrive in recipients' inboxes, while smaller or less profitable entities, like the liberal non-profit liberal advocacy group MoveOn, protested that they were unable to afford the fees and therefore regularly saw their mass mailings filtered as spam.

All three of these scenarios stem from the implementation of a common design principle for accountability: a trade-off between a user's investment in a given online identity and the privileges associated with that identity. This trade-off centers on the idea that individual end-users should be able to decide for themselves how much they wish to invest in their online identities and the size of that investment will then determine the privileges of that identity, within a given application. So a user who invests heavily in an online identity—for instance, by spending money on it, or spending time building up its reputation—will enjoy greater privileges than someone who invests less in another identity for the same application. The basic underlying idea is not unique to cyberspace, indeed it corresponds closely to the notion that "you get what you pay for," but it warrants particular attention when it comes to Internet applications if only because "what you get" (i.e., privileges) and "what you pay" (i.e., investments) can take such a wide variety of different forms. For instance, in the Yelp example the user "investment" takes the form of time spent writing reviews and using the site, while the associated "privileges" consist of features like elite status and prime, highly visible placement of reviews towards the top of a company's page. The user who has invested lots of time in building her profile is rewarded with virtual accolades from Yelp and fellow users, as well as the privilege of having her review prominently featured for other users to see, while the other user who has made only a minimal investment in her Yelp identity is, accordingly, denied the privilege of having her review displayed at all on the company's main page (filtered reviews can still be accessed through a special link at the bottom of Yelp pages). In Second Life, the two users are again distinguished by differing levels of investment in their avatars, however this time it is a financial investment, not a time investment that matters most. The paid account holder is granted the privilege of owning property and building a business on it, as well as deciding which other users (or types of users) will be allowed to access his property, while the avatar associated with the free account has no such privileges due to the low investment. The certified e-mail example presents yet another model of investment-privilege trade-offs; here, the investment is financial—though it is charged per action (or rather, per e-mail) instead of the flat fee levied in Second Life—but the privilege is more akin to that in Yelp, where companies who invest in certifying their e-mail messages are guaranteed the ability to bypass e-mail service filters and reach readers' eyes, while those senders that do not invest as heavily in their e-mail identities run the risk of having their mailings filtered as spam.

7.2.1 How the Trade-off Impacts Accountability

Before discussing other possible implementations of investment-privilege trade-offs, it is important to understand why this design pattern is relevant for increasing the accountability of online identities. The idea that a higher-paying consumer typically receives a better product is, after all, not generally associated with the accountability of either the buyer or the seller, but rather the positive correlation between the price of a good and its quality. This correlation between price and quality is also relevant to many Internet applications, specifically the commercial ones which charge user fees for premium versions of their products, but for Internet application designers the investment-privilege trade-off can also allow for a correlation between price and accountability of an online identity.

Users who pay a higher price for, or make a larger investment in, their online identities significantly reduce the discardability of those identities. They are therefore much more accountable than users who invest less in application identities, maintaining a high degree of discardability and, accordingly, minimal accountability. Thus, user investment in an online identity is closely correlated to the accountability of that identity. Furthermore, in many cases the privilege side of the trade-off can serve as a limiting factor on how much a given user can misbehave within the context of a certain application. For instance, users who have not been granted the privilege of bypassing an e-mail service's spam filters are much more limited in their ability to spam other users than senders whose messages also arrive in recipients' inboxes. Similarly, Second Life avatars who cannot own land or construct new buildings are, in some sense, constrained in how much havoc they can wreak since they are unable to change the physical properties or landscape of the virtual world in as substantive or permanent a fashion as the property-owning avatars. Yelp users whose reviews are routinely filtered or sorted to the bottom of a company's page are much less able to post influential fake reviews than those elite members whose postings are prominently displayed for all viewers to see. In all of these cases, greater privileges correspond to a greater potential for more damaging forms of misbehavior when these privileges are misused.

Thus, the investment-privilege trade-off can also be understood in some sense as an accountability-constraint trade-off, in which the users who can be held most accountable for their actions are the least constrained in their behavior, while those who are less accountable are offered fewer avenues for misbehavior. This is an important perspective to keep in mind when designing investment-privilege trade-off mechanisms both because it explains why this is an effective mechanism for increasing accountability of Internet identities and also because it informs what the most effective forms of investment and privilege are. Investments should not just be something that users are willing and able to spend, they should be something that implies reasonably accurately how discardable an identity is and, by extension, how accountable the associated user is. Similarly, privileges should not just be sufficiently enticing to elicit user payment, they should also be used to relax the constraints placed on a user's behavior. The following sections explore some of these different types of investments and privileges suited to various applications.

7.3 Investments

The challenge of transforming standard discardable, unaccountable online identities into costly, accountable ones motivated the discussion in chapter 2 of different types of user investment in their identities. Investments of either money or time in online identities can sometimes lead to more effective accountability mechanisms when they are incorporated into the model of investment-privilege trade-offs. This model can encompass monetary investment in identities, often in the form of a "freemium" business model, as well as time investments, implemented through reputation systems or initiation periods, and can even be tailored to deal with misbehaviors like spamming which are most easily distinguishable from acceptable application uses based on the frequency of an action (e.g., how many e-mails one sends) rather than the action itself (e.g., sending e-mail).

7.3.1 Financial Investments & Freemium Models

The Second Life and certified e-mail examples both illustrate how financial investments in online identities can be leveraged to determine a user's privilege level. Bulk e-mail senders are given a choice between paying for their mailings to avoid the filters or taking their chances with an unpaid account. Second Life users are similarly given the option between paying for their avatars and enjoying the privileges of owning landing and building up property or, instead, using free avatars to explore and experience the virtual world in a more limited fashion. In this manner, applications can reap the accountability benefits—not to mention profits—of users making financial investments in their identities, while still avoiding some of the main pitfalls of asking users to pay money for accounts by making that payment optional for users who might not be able to afford the fee, or might wish to avoid linking that online identity with a credit card for reasons of privacy or anonymity. Many other online applications feature related two-tier systems in which they offer some basic services for free but require users to pay for "premium" accounts, giving rise to the "freemium" business model.

The term "freemium" was coined in 2006 by venture capitalist Fred Wilson, who conceived of it as an economic strategy to allow Internet services to simultaneously grow their user base and generate revenue (Pujol, 2010). Two common freemium models have arisen since then: feature-limited freemiums and time-limited freemiums. Feature-limited freemium applications provide additional capabilities (or features) to paying customers, while offering a more limited set of services to free account holders. For instance, Google's popular e-mail service, Gmail, operates on a feature-limited freemium model: it is available for free with a limited amount of storage space but users who pay $50 per year for a premium account receive at least 10 GB of mail storage. Time-limited freemium models permit users to access the full version of an application for a free trial period, but then require them to purchase an account if they wish to continue using it. Popular software suites like Microsoft Office and Adobe Photoshop adhere to this model, offering customers 60- and 30-day free trials, respectively.

Economic analysis of freemium models has shown they often increase social welfare over other business models (Niculescu & Wu, 2011). However, little attention has been paid to the accountability implications of the freemium model and the way in which it stratifies users based on the investments they have made in their identities for a given application, limiting the activity of those who are less invested in their online identities and are therefore also less accountable for their online actions.

Since it was designed as a business model rather than a tool for accountability, the freemium approach relies on exchanging premium capabilities for monetary investment. Shifting our focus to embedding accountability in these applications, rather than generating revenue for them, it is possible to imagine variations on this model that also incorporate the methods of time, or energy, investment on the part of the user.

7.3.2 Time Investments & Reputation Systems

Yelp provides us with a good example of how free online identities can be transformed into "costly" ones without charging users any fees but instead relying solely on how much time and energy they've invested in establishing their identities. Yelp extends its version of "premium privileges"—i.e., elite status and prominent placement of reviews—not to users who pay for them but rather to users who have invested sufficient time and energy in building up their Yelp profiles and reputations. A time investment model like Yelp's has some clear advantages over monetary investment strategies: poorer users are not precluded entirely from receiving greater privileges, users are less likely to be discouraged from joining by fees, and users may be more likely to spend more time actually using an application knowing that that time is an investment towards greater status and privileges. Beyond these advantages, however, there are other reasons why a time investment system is a more effective tool for accountability than charging monetary fees would be for applications like Yelp.

The form of misbehavior Yelp is most concerned about damaging its site is fake or intentionally misleading reviews from paid authors or business owners themselves. In this case, charging users a fee for the privilege of placing their review more prominently could actually lead to more of this sort of misbehavior, since many businesses have already demonstrated that they are quite ready and willing to spend money on positive Yelp reviews by paying other people to write them. One can easily imagine a scenario in which charging users more money to put their review higher up on a business' page would mean that most of the reviews listed first were posted by the business itself, or people it hired, rather than by legitimate users, who would probably be less willing to pay for the privilege of reviewing a restaurant. Therefore, in applications like Yelp where the anticipated malicious activity is likely to be driven by well-funded companies, charging fees for user privileges is much less likely to drive accountability than time investment models, and could even have the opposite effect of encouraging more rampant misbehavior. Paid review site Angie's List addresses this risk by collecting users' real names and providing these names to businesses who wish to respond to their reviewers, reducing (though by no means eliminating) the risk of fake reviews.

Malicious activity on the Internet runs the gamut from harmless irritations to more serious forms of harassment, impersonation, and even criminal activity. Developing an effective accountability mechanism requires identifying not just what kinds of misbehavior a given application might be subject to, but also who the likely perpetrators would be and what the best means of discouraging them might be, whether it's charging them money, or requiring that they spend extended periods of time developing online identity reputations, or something else entirely. Therefore, applications like Second Life that are mostly plagued by individual griefers who are unlikely to want to spend large sums of money on their online identities can use paid accounts to greater effect than applications like Yelp which fear misuse by well-funded companies and are therefore reliant on time-investment reputation mechanisms.

Notably, Yelp also has a more typical "freemium" business model for companies that are reviewed on its site. Company owners may purchase paid business accounts from Yelp tied to their listing on the site, in order to make use of Yelp's search engine marketing product that lets their listing appear more prominently in an area above usual search results. This marketing tool allows Yelp to profit from paid accounts without undermining the integrity of the user reviews. Different kinds of investments and privileges are not necessarily mutually exclusive. It is often possible to combine elements of financial and time investment in a ways that can both promote greater accountability and offer more choice to users.

7.3.3 Authentication Investments

One alternative to asking users to invest time or money in their online identities is asking them to invest an indicator of their real identities by authenticating themselves to the application operator. This type of authentication investment could provide an insurance policy of sorts to the application owners who know that when users misbehave it will be possible to identify the responsible parties and hold them accountable in the real world. Thus, users who were willing to invest in their online accounts by authenticating their real-world identities—most likely by means of submitting credit card information—could be granted additional capabilities or privileges. If the authenticated account holders abused these privileges, it would then be possible to charge them fines for damages inflicted using the submitted credit card information, or hold them accountable under the judicial and punitive mechanisms of the physical world, or, in less serious cases, simply prevent them from opening any future accounts using the same credentials. Of course, a user could have multiple credit cards and therefore be able to open multiple accounts, but this would still have the effect of severely limiting the repeated creation of easily discardable identities.

Authentication investments are a form of conditional anonymity, in which Internet users' real-world identities are encrypted, kept secret, or held in escrow by a trusted third party unless they engage in malicious activity. Conditional anonymity schemes as a tool for accountability are explored in greater detail in chapter 8, but it is worth noting that they can also play a role in investment-privilege trade-off mechanisms.

7.3.4 Investing in Actions vs. Identities

One final note on the nature of investments in designing investment-privilege tradeoffs is that they can be tied either to the cost of a single identity or to a specific action or privilege. For instance, Second Life users can invest money in their avatars by paying Linden Lab a set monthly fee, but certified e-mail senders were required to pay per e-mail instead of a flat fee. This distinction matters because some forms of misbehavior, most notably spamming, are most easily distinguishable from normal or acceptable application use by their frequency rather than their actual content. If a Second Lifer defaces the John Edwards campaign virtual headquarters, that action could be considered malicious by the application and its other users not because of how often it was done but because of *what* was done. By contrast, a spammer would be most unlikely to send only one unsolicited commercial e-mail. In other words, we identify misbehaving avatars in Second Life based on *what* they do, but we are more likely to detect e-mail spammers based on *how often* they do what they do.

Identifying the nature of the malicious activities that may arise in an application is crucial to designing appropriate accountability mechanisms to prevent them. Activities like spamming that can be distinguished from regular uses of an application, like e-mail, primarily based on the frequency with which they occur are better suited to investment per action, rather than a single, set investment in individual identities. By contrast, in applications where an individual action can be malicious on its own, it makes more sense to require users to invest in their identities as a whole and use that investment to hold them accountable for what they do, rather than how often they do it. It is also possible to combine these two forms of investment and allow users to invest in both their online identities as well as the individual, rate-limited actions they undertake using these identities. In some cases, the level of investment required for the latter may depend on how much users have invested in their online identities overall and how well-established they are in the context of a specific application.

7.4 Privileges

Perhaps even more than investments, user privileges can vary enormously from application to application. In Second Life, users pay for the privilege of owning property, in certified e-mail schemes the privilege on offer is bypassing spam filters, in Yelp it is the prominence of posted reviews, in Gmail it is expanded storage capacity. Given the great diversity of Internet applications it can be difficult to classify the specific categories of privileges they present, but it is possible to derive some general characteristics and principles of common application privileges that can be used as counterparts to users' investments in their online identities. These different types of privileges include: prioritization, prestige, expanded capability, influence over oneself, and influence of other users. As with investments, different privileges can also be combined and used to reinforce each other within the same application.

The most appropriate privileges to promote accountability—like the most appropriate investments—often depend on the form and function of a given application. One of the key questions when evaluating appropriate privileges for an application is to what extent withholding that privilege from users prevents them from misbehaving, since ideally a designer would like less accountable users to have less capacity to cause trouble. This is a separate issue from what types of privilege will be most likely to entice users to invest more in their identities, but it is a crucial element of an effective investment-privilege trade-off: that the investment not only be used to promote accountability but the privileges also be leveraged to constrain the boundaries of less-invested users' capacity for malicious activity.

7.4.1 Prioritization

Yelp users and certified e-mail senders both rely on their identity investment to bypass spam filters and receive priority over the information posted or sent out by competing users who have invested less in their online identities. In other words, both applications reward users who are more heavily invested in their identities by prioritizing their online activity over that of other users. For Yelp and certified e-mail this prioritization takes the form of prominent placement of postings and e-mails to be easily viewed by others, but prioritization can also be based on bypassing other constraining mechanisms besides spam filters, for instance, waiting periods for an online service or moderation by site administrators.

Granting prioritization privileges only to users who are more heavily invested in their online identities can be an effective means of constraining misbehavior by users who are less invested in their identities. For instance, Yelp's prominent placement of reviews by users who have invested more in their reputations constrains the ability of less invested users to post false reviews that are likely to be seen or read by many other users. Similarly, online forums and commenting applications that allow registered users' comments to be posted automatically, without moderation, prevent users who were not willing to invest in the registration process from being able to post spam, offensive content, or otherwise problematic messages for others to see, since these posts can be deleted by moderators prior to publication. Thus, for applications like Yelp and Wikipedia, which focus on providing relevant and reliable user-generated content, prioritization of postings or submissions by users who are more heavily invested in their online identities can be a highly effective means of constraining the ability of less invested users to undermine the overall quality of the information provided.

7.4.2 Prestige

Elite Yelp status does more than convey prioritization privileges on selected users, it also provides the less tangible reward of user prestige. Users who are awarded elite status don't just get to post their reviews prominently, they also receive annual virtual "badges" announcing their status which are displayed on their profiles and next to their reviews. Similar systems of online merit badges exist across numerous different applications, including TripAdvisor, FourSquare, the Huffington Post, and Wikipedia, whose most diligent contributors are rewarded with virtual barnstars, shown in Figure 7-1.

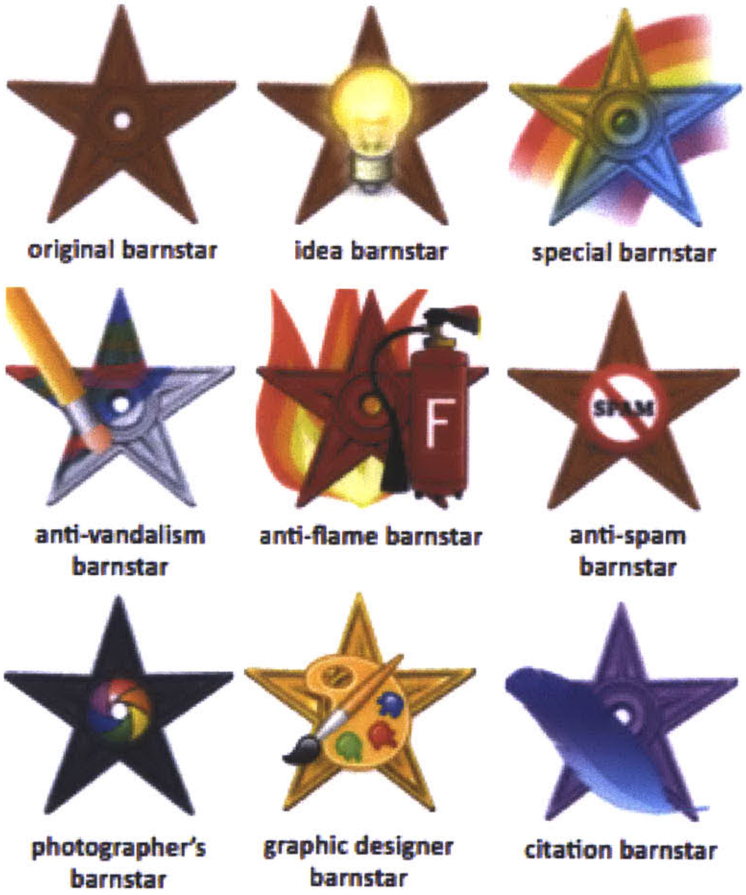


Figure 7-1: Wikipedia rewards its most dedicated users with barnstars that identify their contributions to the site.

Antin and Churchill (2011) note:

> Badges can be motivating as status symbols. Badges advertise ones achievements and communicate ones past accomplishments without explicit bragging ... Badges also provide personal affirmation in that they serve as reminders of past achievements much like trophies on a mantelpiece. They mark significant milestones and provide evidence of past successes.

Recognition by the application and fellow users, in other words, can be a major incentive for encouraging greater user investment in online activities. Similarly, high rankings on application leader boards as a top commenter or reviewer can serve as a psychological form of privilege for users who have made significant investments in their online identities and are gratified by receiving some public, application-wide acknowledgment for those investments (Sproull, 2011).

In some cases, this prestige may come with other, more concrete, perks. For instance, Amazon's Vine Voices provides members not just with a virtual icon depicting their status but also with free products to reviews. Similarly, according to Yelp's website, "Yelp Elites get a nifty annual badge on their Yelp profiles, and have the opportunity to get invited to exclusive local events and parties hosted by Yelp." So for both Amazon Vine Voices and Yelp elite users, the status badges are augmented by other privileges like free products and exclusive social gatherings. Combining prestige privileges with more tangible benefits can undoubtedly strengthen their appeal, but badges, status symbols and other public indicators of user prestige can play an important role in investment-privilege trade-offs, serving not only as a signal of a user's status to others but also as a mental boost and symbol of honor for the user. Brennan and Pettit (2004) note, "the desire for esteem will play a significant role in stabilising decent communication practices and supporting the operation of other social norms on the Internet."

7.4.3 Expanded Capabilities

Another form of privilege is simply to expand, rather than prioritize, a user's capabilities within the context of an Internet application. Paying for extra storage space from e-mail applications like GMail or content-sharing applications like Dropbox is one common type of expanded capability privilege. Examples of expanded capabilities could also include allowing users who are more invested in their identities to perform an action or use a service more often than others who have made smaller investments in their online identities. For instance, an e-mail application that limited the number of e-mails its users sent to prevent spamming might allow users who were sufficiently invested in their accounts to send unlimited messages. Alternatively, a service that was only offered for a limited period of time or a limited number of trials to less invested users could be made available for longer periods or unlimited use to more heavily invested users. These types of privileges are often seen in the time-limited and feature-limited freemium models discussed previously. For applications where malicious activity requires repeated, extensive or especially frequent use (e.g., spamming, harassment), denying less-invested users these privileges can be a useful way of preventing misbehavior on the part of less accountable users.

Users of applications that offer expanded capability privileges invest in their accounts in exchange for expansion, or extension, of the privileges they already possessed as less invested users. Similarly, prioritization focuses on allowing users to do mostly the same things they were able to do with identities that they were less invested in but now their actions are prioritized over those of many other users. In other words, neither prioritization nor expanded capability privileges actually grant users substantively new abilities, instead they extend and expand on the abilities given to less invested users. Another option is to extend new abilities to users who have invested more in their identities, rather than merely expanding or prioritizing the ones they already had. These new privileges generally offer users either the ability to exert greater influence over their own personal experience with the application or the ability to exert greater influence over the experience of others using the application.

7.4.4 New Capabilities: Influence over Oneself & Other Users

The most straightforward set of new capabilities that users can be granted in exchange for investing in their identities are those that allow them to alter the nature of their own experience within an application. For instance, users who pay for ad-free accounts are given the privilege to improve their own personal application experience, without affecting the experience or behavior of any other users. In Second Life, however, the privilege of owning property affects not just the experience of the user who can own land but also the experience of other avatars within the application who may see that property and be granted or denied access to it by the owner. Some privileges are even more directed at empowering individuals to influence the behavior of other users: Wikipedia's system of administrators, bureaucrats, and other access levels for users is almost exclusively concerned with granting more invested users privileges to monitor and moderate the actions of their fellow users. For instance, Wikipedia administrators are given privileges to block and unblock other users, as well as to delete or protect pages on the site.

Additional privileges, as opposed to prioritized or expanded ones, can therefore be used to build new user-based, hierarchical accountability mechanisms within applications in which some users are given privileges to hold others accountable. In other words, granting new privileges to certain users who are heavily invested in their identities can allow them to hold other, less invested users accountable, enabling more complex participatory governance mechanisms. In community-based applications, where many different users engage in group settings, these privileges that allow some users to have influence over the behavior of others are generally much more effective for accountability purposes than additional privileges aimed only at allowing individuals to influence their own experience within the application. However, applications focused more on the formation of bilateral relationships and interactions between individual users can benefit from privileges that allow some users to control their personal experience more carefully, by allowing them either to block other users or to view more information about the other users with whom they interact. In general, however, these sorts of personal end-user controls are not necessarily ideal for restricting to more invested users, since less invested users could often benefit just as much from them and granting them these privileges would oftentimes not enable less accountable users to engage in greater malicious activity. In other words, these privileges may be a useful element of an investment-privilege trade-off meant to encourage greater investment but offering them only to more invested users does not necessarily increase the level of accountability and constraints on misbehavior of the application, as compared to providing them to all users.

7.5 Trade-offs & End-User Customization

Allowing users to decide for themselves how much they want to invest in their online identities and then adjusting their account privileges and capabilities accordingly can give rise to accountability mechanisms that are not just tailored to specific applications, according to their function, but also personalized by specific end-users, according to their preferences. Design decisions have a tremendous impact on the choices and customization available to users. Clark and Blumenthal (2011, p. 382) explain that "application design and modularity can enhance or reduce options for user choice. Different designers will have different motivations to offer or constrain choice, and thus control the degree to which a user can make personal decisions about trust within specific applications."

Designing for personalization of identity schemes goes beyond the idea that different applications should have different accountability mechanisms to the possibility of different users being subject to different accountability mechanisms, even within the same application. The Center for Strategic and International Studies report on *Securing Cyberspace for the 44th Presidency* advocates this approach, stating: "consumers should have choices about the authentication they use and be able to tailor the identity information they exchange to provide the minimum needed for a transaction. Allowing consumers to choose and to tailor the information they provide will enhance both security and privacy" (Langevin et al., 2008, p. 64). One means of achieving this goal is allowing application users to make trade-offs between how well-established their identities are and how many privileges or capabilities are associated with those identities. Users who are willing to invest more time or money in their identities are granted access to a wider variety of activities in a given application, while users with newer, or less well established identities may be unable to exercise as many privileges, or rate-limited in their actions, or required to pay some further fee to enjoy the full range of the applications capabilities.

The value of the investment-privilege trade-offs for accountability described in this chapter lies largely in the freedom they afford users to decide, for each application they use, how much they value accountability, anonymity, and the application privileges on offer, and then tailor their various online identities to these preferences. In a sense, trade-offs of this nature permit individuals to determine for themselves where they wish to fall on the anonymity-accountability axes. Furthermore, users can adjust this trade-off from one application to another, allowing them even greater freedom to use the Internet for a myriad of different purposes. For the application designer and operator, these systems have the benefit of liberating them from the burden of implementing a one-size-fits all accountability scheme for all of their users and the associated headache of trying to determine which such mechanism which will attract the most users while still affording the desired degrees of accountability and anonymity. Thus, these trade-offs can provide all involved parties significant flexibility and freedom in tailoring online identities and accountability mechanisms.

These trade-off accountability mechanisms are also primarily self-regulating and inward-facing, meaning they are largely independent of Internet jurisdiction disputes and involvement of national governments. For Internet applications with a global

Table 7.1: Examples of Design-Privilege Trade-offs in Online Applications

Application	Investment	Privileges
Yelp	**Time investment:** posting numerous, thoughtful reviews and acquiring complimentary ratings from other users	**Prioritization:** reviews posted at the top of business pages **Prestige:** elite user status and badge
Amazon	**Time investment:** writing regular product reviews and receiving large volumes of high rated by other users	**Prestige:** membership in Amazon Vine program **New capabilities:** receiving free, advance products to review
Second Life	**Financial investment:** paying a monthly fee to Linden Lab **Authentication investment:** submitting credit card information, with no fees charged	**New capabilities:** owning land and restricting who can access that property **Prestige:** status as a verified user is publicized to all other avatars in user's profile
Wikipedia	**Time investment:** registering a user account, contributing thoughtful, useful entries as well as editing and correcting existing or inaccurate entries	**Prestige:** barnstars for exceptional contributions **New capabilities:** moderating and locking entries, blocking and unblocking users
GMail	**Financial investment:** annual or monthly fee	**Expanded capabilities:** more storage space
GoodMail	**Financial investment per action:** bulk e-mail senders pay a per-message fee to a certification service and its partner e-mail providers	**Prioritization:** e-mails bypass spam filters of selected e-mail providers **Prestige:** e-mails flagged as "certified" in recipients' inboxes

user base, holding individual users accountable for their actions in court can be a challenging and time-consuming endeavor. Even if an application owner has reliable evidence of the real-world identity of one of its users, it will not necessarily be possible to use that real identity to hold the responsible party accountable if the offending user is located in a different jurisdiction from the application owner, or a jurisdiction where the behavior does not warrant disciplinary action. In this respect, investment-privilege trade-offs that encompass elements of both anonymity and accountability and rely on private actors like application owners and credit card companies for enforcement can sometimes be even more effective than schemes that identify the real world identity associated with every online user but rely on legal enforcement administered by numerous different jurisdictions.

Designing an application with the proper mix of options for investing in online identities and rewarding those investments with appropriate privileges that relax the constraints placed on user behavior can have many advantages. Defining these trade-offs can enable application designers and operators to determine the space of accountability and anonymity preferences they believe to be appropriate for their application. Giving end-users the power to decide how much to invest in their identities and how many privileges to acquire or sacrifice can, in turn, empower individual users to figure out where they want their online identity for this application to fall within the space defined by the designer. Finally, leveraging user investment to increase accountability of online identities and constraining user behavior through granting and denying privileges based on these investments may even, in some cases, help application operators avoid having to grapple with complicated issues of jurisdictional and legal enforcement.

Chapter 8

Design Patterns for Conditionally Anonymous Identities

> Those who use the Internet to recruit terrorists or distribute stolen intellectual property cannot divorce their online actions from their real world identities. But these challenges must not become an excuse for governments to systematically violate the rights and privacy of those who use the Internet for peaceful political purposes.
>
> —Secretary of State Hillary Clinton, "Remarks on Internet Freedom," January 21, 2010

Allowing users to weigh how much they want to invest in their online identities against how many privileges they want associated with those identities can be a powerful means of holding anonymous identities accountable. However, the trade-off mechanisms described in the previous chapter are not always sufficient or appropriate for dealing with every type of misbehavior that arises in Internet applications. Applications whose users engage in criminal activity, for instance, may require more rigorous means of holding those users accountable in the legal systems of the real world beyond simply tarnishing their online reputations or diminishing their privileges within the application. In these cases, one approach to reconciling anonymity and accountability in online identity schemes is to implement conditional anonymity mechanisms that protect the anonymity of users' online identities so long as those users do not violate certain rules or terms of service. Conditional anonymity schemes involve users providing some form of identity authentication, either to an application's operator or to a trusted third party, when creating new online identities. This authentication information is not used to trace the user's real identity, however, except under specific conditions of misbehavior.

By definition, these conditional anonymity mechanisms do not provide users with total anonymity but, at least for some applications, they may be able to offer sufficient privacy protections to satisfy users while still affording application owners some effective tools for enforcing accountability. Such schemes are beneficial not just for application operators, however. Recall that when Linden Lab announced that it would no longer require all Second Life users to verify their identities by submitting credit card information, many existing users protested the decision and circulated a petition to try to stop the change, due to their worries that it would noticeably degrade behavior and the quality of user interaction within the community.

8.1 Internal & External Accountability

Undoubtedly, conditional anonymity mechanisms are inappropriate for certain types of applications, especially those that require the highest degree of anonymity, but these schemes can still play an important role in improving user accountability for some applications. In particular, when an application's internal mechanisms do not suffice to hold users accountable, it may be necessary to turn to external accountability mechanisms administered by the outside environment, or society. The investment-privilege trade-off mechanisms discussed in chapter 7 are primarily a means of creating internal accountability, in which users are identifiable (often by a pseudonym) within the context of a particular application and can be held responsible for their actions within that context, according to the norms and policies of that application. These internal mechanisms are largely independent of national borders and jurisdiction, which is often–but not always—an advantage. In cases where it may be desirable for a user to face legal action, for instance, the investment-privilege trade-offs can be problematic since they do not usually lend themselves easily to imposing external accountability. External accountability requires that users be identifiable in the real world so that they can be held responsible for their actions according to the laws and regulations of their external environment, outside the context of the specific application in which those actions took place (Farkas et al., 2002).

Internal accountability can be used to regulate the behavior of virtual entities and penalize online identities, while external accountability mechanisms are used to hold real-world users responsible for their online actions. External accountability therefore requires being able to link online identities with the people who create and control them. Just because it is possible to perform this attribution, however, does not mean that all users need automatically be identified in this manner. Instead, it may be possible for applications to guarantee that its users will retain their anonymity under certain conditions of good behavior. This notion of conditional anonymity is related to the idea of fair cryptosystems, introduced by Micali (1993). Fair cryptosystems involve users breaking their private decryption keys into five pieces and entrusting each piece to a trusted entity, for instance, a federal judge. Without all five pieces, it is impossible to retrieve a user's private key and decrypt their communications, but when all five trusted parties agree it is necessary (for instance, when a court order is issued) they can recover a user's key. Micali (1993) notes that such systems are both "unabusing" because the privacy of law-abiding users cannot be compromised and "unabusable" because users who break the law do not enjoy any privacy. Similarly, conditional anonymity schemes are intended to ensure that the anonymity of well-behaved users is strongly protected while the real identities of misbehaving users can be accurately determined in order to hold them accountable.

In this chapter, three different design patterns for implementing application-layer conditional anonymity schemes are described and evaluated based on the advantages and disadvantages of each in various types of applications. First, we look at applications that choose to access and store their users' real identities themselves—usually by means of authenticating credit card information—but do not associate those real identities with their online counterparts in any public manner, except in the case of extenuating circumstances. Second, we turn our attention to encrypted identity mechanisms that allow users to encrypt their real identities in such a manner that they can never be accessed by applications, or can only be accessed in very specific situations dictated by the encryption algorithm. Finally, we examine methods of identity escrow and discuss the challenges of enforcement and jurisdiction that often arise when dealing with conditionally anonymous identities.

8.2 Authentication By Applications

An application can be designed to give its operator access to users' real identities in a variety of ways, but perhaps the most straightforward method is for the application simply to collect a record of the real identity associated with each of its users. Of course, it does not suffice just to ask every new user for a real identity, the application needs some way to authenticate that identity for it to be useful. In the real world, authentication can be a time- and labor-intensive process for both the person being authenticated, who often has to submit multiple forms of identification (driver's license, passport, etc.), and also the authenticator, who must collect the requisite documents, verify their authenticity, record any necessary information, and then return them. Clearly, it would be impractical to model mechanisms for Internet applications on the authentication procedures we use for real-world activities like opening bank accounts or obtaining visas. Instead, Internet applications generally rely on credit card companies to authenticate the identities of their users. Prior to 2006, for instance, when all Second Life users, even those who held unpaid accounts, were required to verify their identities, new users had to submit a credit card number and the associated billing information to Linden Lab in order to create an avatar. Credit card companies then authenticated the user's real identity for Linden Lab, which could store that information for future use, as insurance against misbehavior. Meanwhile, the new avatar could enjoy complete anonymity in Second Life so long as Linden Lab did not reveal his true identity to any other users.

Mechanisms like these provide application users with anonymity that is conditional on the judgment and discretion of the application administrator or operator, which possesses the necessary information to trace an online identity back to a specific person but exercises this ability only when deemed necessary. This provides application operators with a greater range of options for holding their users accountable but can also have the disadvantage of discouraging many users from signing up, either because they wish for more thorough anonymity or because they do not want to disclose their credit card information. As Linden Lab discovered when it removed the identity verification requirement and saw its membership increase fourfold, just asking people to submit a credit card number—even if they are not going to be charged any fee—can have the effect of deterring many users.

8.2.1 Credit Card Authentication

Credit card validation is not the only possible means of identity verification available to online applications, but it is a particularly popular method since it allows application operators to pass on the burden of authentication to the credit card companies. However, credit card authentication can place other burdens on application owners, who must pay credit card agencies to perform the authentication process and also securely collect, and sometimes store, users' credit card numbers and other personal information. Secure storage of users' identifying information can be a particular challenge for applications. In September 2006, for instance, one of Linden Lab's databases containing unencrypted customer information was breached, allowing the hacker to access users' names and addresses (Veneziani, 2006). Smaller or newer applications without the necessary resources to devote to data security may be particularly wary of authenticating users and opening themselves up to the possibility of data breaches.

For commercial applications that sell items or services to their users, the burdens associated with collecting and storing identifying information are necessary costs regardless of issues of accountability. However, for applications that wish to use conditional anonymity schemes purely for the purposes of being able to hold users more accountable, instead of for commercial transactions, the costs of credit card verification can be prohibitive. The U.S. Supreme Court addressed these costs in its ruling on the1997 case *Reno v. American Civil Liberties Union* concerning the constitutionality of the Communications Decency Act, which required website owners to shield minors from viewing inappropriate content by using information like users' credit card numbers to verify their ages. The Court noted that, "Using credit card possession as a surrogate for proof of age would impose costs on noncommercial Web sites that would require many of them to shut down" (*Reno v. American Civil Liberties Union*, 1997). Similarly, using credit card verification for accountability purposes may not be feasible for many non-commercial applications. Additionally, the Court also pointed out, requiring users to submit credit card information even for non-commercial purposes blocks access to everyone who does not have a credit card.

8.2.2 Identity Certificates & Credentials

There are alternatives to credit card-based authentication schemes. Rather than providing applications with credit card information, users could instead give application owners signed identity certificates that indicate their real identities. For example, the United States government's *National Strategy for Trusted Identities in Cyberspace* (2011) envisions an "identity ecosystem" in which users have the option of maintaining different, interoperable certificates, or identity credentials, from various private and government actors. Relying on non-credit card identity credentials might reduce the risks of data breaches for applications and their users. If an application's security is breached and all that is revealed are the real identities of its users, rather than their credit card numbers, this would at the very least be a less financially damaging consequence for users. Reducing the incentives for data breaches and the associated risks could liberate application designers and operators from some of the data security concerns and other costs associated with storing users' credit card numbers. At the same time, users may also feel more comfortable sharing non-credit card identity credentials with these applications, and users who do not have credit cards would not be excluded from accessing applications under these conditions.

8.2.3 The Role of Application Operators

Regardless of whether they rely on credit card validation or other identity credentials, applications that authenticate their users' real identities retain the privilege—and the responsibility—of deciding for themselves under what circumstances an online identity should be traced back to its real-world counterpart and how that real identity should be used to hold a person accountable. In other words, these schemes render a user's anonymity conditional based on the subjective opinions and unilateral decisions of the application operator. Depending on how much users know about and trust a given application, this may or may not be a desirable situation from their perspective. There are methods of mitigating or eliminating this subjectivity, however, including the alternative conditionally anonymous systems discussed in later sections of this chapter that involve identity encryption or relying on a trusted third party, rather than the unilateral decisions of a single application, to protect real identities.

The applications best suited to directly storing the real identities of their users are commercial ones which must bear the costs of credit card number collection and verification anyway, in order to conduct business. In this case, the accountability benefits of having a real identity associated with every online account are essentially a side effect or automatic added perk of the application's business practices. Concerns about subjective judgments by application operators regarding how to use these identities are still an issue, but since users have already indicated they trust the application enough to use it for commercial transactions, it does not seem unreasonable to assume they may also be comfortable trusting the application's accountability mechanisms. The application operator, in turn, is given the flexibility to decide when it wants to leverage users' real identities for accountability purposes and how. For instance, some application operators may choose to punish misbehaving users by blocking or deleting their account and then banning another account from being created using the same credit card or another credit card with the same billing information, in order to reduce identity discardability. Another option might be for applications to hold their users accountable for misbehavior by charging fines to their stored credit card information. In the case of criminal activity, application owners could instead choose to pass on information about users' authenticated identities to law enforcement agencies.

Application operators are also given great freedom under this system to define what they believe constitutes inappropriate behavior within the context of their applications and can revise and update that definition constantly, as they witness new forms of malicious activity. In other words, the application owner enjoys maximal flexibility in determining what types of behavior merit which kinds of punishment, but users will potentially have to deal with constantly changing and highly subjective decisions about when and how their real identities will be used. Concerns about how real identities may be misused under these relatively lax constraints have given rise to some research on algorithms for so-called "unconditional anonymity," in which users' anonymity is cryptographically protected in all except very specific, pre-determined circumstances (Hayes, 1990).

8.3 Encrypting Real Identities & Unconditional Anonymity

Research on encrypted identity and unconditional anonymity schemes for online identities originated in large part from the growing prevalence of Internet commerce and the inability of users to make anonymous purchases online since users who make online purchases with their credit cards have "no protection against surveillance" (Chaum, Fiat, & Naor, 1990). As a means of increasing consumer privacy in this area, Chaum (1983) developed a system of blind signatures, in which messages are encrypted, or "blinded," before being digitally signed so that the signer does not know the content.

8.3.1 Digital Cash

Applied to digital payment systems, blind signatures can allow banks to authorize (or "sign") payments after the identity of the sender had been blinded, allowing for the development of untraceable, or anonymous, payment systems. Such payment systems are often considered the electronic equivalent of cash in that it is quite difficult to trace paper money back to its spender, but their digital implementation presents some unique challenges. "Paper cash is considered to have a significant advantage over credit cards with respect to privacy," Chaum et al. (1990, p. 319) point out, "but what is to prevent anyone from making several copies of an electronic coin and using them at different shops?" To prevent fraudulent spending, several proposed systems of digital cash encrypt users' identity credentials in such a way that they cannot be decrypted unless a user tries to spend the same cash tokens multiple times. In other words, users' anonymity is guaranteed unless they engage in a very specific and well-defined form of misbehavior (in this case, fraud).

However, even as they provide valuable privacy protections, blind signatures afford users a potentially dangerous degree of anonymity. Blind signature schemes can detect and prevent specific types of fraudulent behavior, but they do not allow for the linking of real identities to specific actions which can, in some cases, enable anonymity protection for other types of criminal behaviors (Solms & Naccache, 1992). For instance, blackmailers could use the anonymity protections afforded by blind signatures to perpetrate "perfect" (or untraceable) crimes by forcing a victim to "anonymously withdraw digital money from his account, acting as an intermediary between the blackmailer and the bank" (Stadler, Piveteau, & Camenisch, 1995). Since the blind signature scheme would prevent the ransom from being identified later on, it would then be impossible to catch the blackmailer. Thus, while protecting real identities with blind signature encryption can be an extremely powerful means of protecting anonymity, that protection can also be vulnerable to exploitation and misuse.

To help deal with situations like these, Stadler et al. (1995) propose a mechanism of "fair blind signatures" which are similar to Chaum's blind signatures but also include a "link-recovery" protocol which allow the signer to obtain the message he signed from a trusted third party, or judge. Introducing the potential for anonymity revocation into these systems allows for greater flexibility when dealing with a range of different kinds of malicious activity, but also creates the potential for abuse and even the possibility of spurring bad actors to commit worse, or more complex, crimes as they seek to evade the revocation protocols. Davida, Frankel, Tsiounis, and Yung (1997) describe seven requirements for implementing anonymity controlled identity mechanisms for digital cash, including:

1. Anonymity for legitimate users who do not engage in prohibited activities.
2. Revocation of anonymity by a trusted party when necessary, as judged by that party.
3. Inability of the trusted party to forge coins or impersonate users, i.e. separation of power between the trusted third party and the coin issuing agency.
4. Impossibility of the bank framing users for malicious activity, even in collaboration with the trusted third party.
5. Selective revocation of anonymity so that only a specific, targeted transaction is de-anonymized while all other transactions, even those involving the same user, remain anonymous.
6. Anonymity revocation should be efficient and create minimal burdens for all involved parties.
7. Anonymity revocation should not, directly or indirectly, spur more serious crimes than the ones it prevents.

The number of requirements here hints at the complications and risks associated with such revocable anonymity schemes. Still, such methods have been developed using fair blind signatures, but they have not gained widespread popularity or deployment, perhaps due in part to the failure of digital cash to attain a strong foothold in the world of online commerce (Qiu, Chen, & Gu, 2002).

8.3.2 Applications Beyond Payment Systems

Using encryption to protect identities is not only relevant for online cash systems, however. Afanasyev et al. (2011) apply a system of identity encryption to the network layer of the Internet with their proposed "privacy-preserving network forensics" in which every packet contains a cryptographic signature that, under suitable circumstances, can be used to identify unambiguously the physical machine that sent the packet. "Without a plausible threat of accountability, the normal social processes that disincentivize criminal behavior cannot function," the researchers note. "We suggest modifying the Internet architecture to proactively enable network forensics while preserving the privacy of network participants under normal circumstances." The privacy-preserving feature relies on the notion that "absent express authorization" the signatures would not reveal any identifying information, however the authors' vagueness regarding what precisely would constitute proper authorization tempers this guarantee to some degree and raising similar concerns about subjectivity to those discussed above. Similar unconditional anonymity methods could be applied at the application layer by attaching cryptographic signatures to users' online identities rather than individual packets, though the definition of what specific authorization would be needed to reveal these identities would remain a critical component of any such system. Blind signatures, in contrast to these more subjective mechanisms, provide much more rigorously defined conditions of anonymity but are also more limited in their applications.

Micali's notion of fair cryptosystems, in which multiple actors must cooperate to retrieve a user's encryption key, also suggests another possible extension of Chaum's blind signatures that could apply to a broader range of Internet applications than just digital cash. If users entrust their real identities to some trusted third party, that trusted entity could potentially sign an encrypted binding between the user's real identity and a pseudonymous, virtual identity for the user to use within the context of a specific application. If decrypting this binding required the cooperation of both the application operator and the trusted third party, in the same way that fair cryptosystems require the cooperation of multiple parties to retrieve a private key, it would be impossible for either the application or the third party to unilaterally link an online identity to a real person. This way, applications could be reassured that identifying their users would be possible with the help of a trusted third party, even without that trusted party having immediate and independent access to the online identities of all of its clients.

Though such a system would retain some of the subjectivity inherent in application authentication schemes, by requiring the cooperation of two, independent parties it could reduce the risks of misuse, security breaches, or corruption. It would also mean that the trusted third parties could not immediately identify whether a misbehaving user has multiple identities within a certain application, or across different applications, unless each of these identities individually engaged in malicious activity. This could, in turn, make it more difficult for applications to ban problematic users altogether forcing them, instead, to terminate misbehaving identities one at a time. Additionally, trusted third parties who stored real identities for multiple applications would be unable to alert all the different applications used by an individual malicious actor when that person caused trouble in one application. In other words, it can be very difficult under this system for applications or third parties to link together all of a given user's online identities. Different conditional anonymity mechanisms address identity linkability differently, but it is a particularly salient issue for other systems which involve having a trusted third party store users' real identities, rather than giving them directly to an application. This type of conditional anonymity mechanism is known as identity escrow.

8.4 Identity Escrow

When end-users do not want to trust an application with their real identities or applications do not want to bear the responsibility of storing and protecting those identities it may be possible to implement conditional anonymity by means of identity escrow. Identity escrow mechanisms work by users entrusting their real identities to some third party, which in turn guarantees to the application operator that it can determine the user's identity should the need arise. In this fashion, applications are not responsible for storing users' real identities and users can be reassured that the applications they use will not be able to access their real identities except in cases when the trusted escrow agent agrees to divulge them (Kilian & Petrank, 1997). Identity escrow mechanisms therefore "allow users of a service to remain anonymous, while providing the possibility that the service owner can break the anonymity in exceptional circumstances, such as to assist in a criminal investigation" (Mukhamedov & Ryan, 2005). In some cases, escrow schemes can alleviate the burdens on both users and application owners while still ensuring an acceptable degree of both anonymity and accountability.

8.4.1 Escrow Requirements

Users may establish "escrow requirements" with their trusted third parties, outlining well-defined circumstances under which their identities will be revealed and guaranteeing that unless those conditions are met they will be able to maintain their anonymity. Such agreements bear some resemblance to Chaum's digital cash mechanism, which guarantees that users' identities will only be accessible if they act in very specific ways that allow for the decryption of their identities, however, there are two notable differences. Purely encryption-based approaches to conditional anonymity can only be used to protect against behaviors that can be leveraged to automatically trigger decryption (e.g., repeat spending of digital coins). Escrow agents can set more varied and complicated conditions, such as releasing information of interest to ongoing criminal investigations, that could be more difficult to hardwire into encryption schemes that do not involve a third party. Second, hardwired encryption-enabled anonymity schemes offer little flexibility for adjusting the conditions of anonymity or relaxing them in exceptional circumstances. This is by design—the point of Chaum's digital cash system is that it eliminates the subjectivity and flexibility of parties to determine what circumstances merit tracing a user's real identity. However, in some cases there may be reasons why it is impossible or inadvisable to so strictly delimit the circumstances in which a user can be linked to his real identity. In these instances, allowing an identity escrow agent to exercise its judgment in determining when it is appropriate to release a user's identity can be a suitable compromise. Ideally, this can allow for some flexibility in dealing with unexpected situations but still provide end-users with some reasonable assurance of privacy, so long as the third party is trusted by both the user and the application to make these decisions.

Applications that rely on identity escrow mechanisms, like end-users, can also negotiate arrangements with their trusted escrow agents to hold individual users accountable for specific sets of behaviors or circumstances and in a variety of different ways. Some applications may simply want the escrow agent to give it the name and identifying information of the offending user. In other cases, if the application's priority is ensuring financial liability of its users, the escrow agent may have secured a monetary bond from the user and simply pay the application owner that sum when certain conditions are met, without even needing to reveal the user's identity to the application. An alternative might be for the escrow agent to act as an insurance carrier or a representative on behalf of the user for resolving disputes with the application (Aura & Ellison, 2000). This flexibility with regard to not only the circumstances under which an escrow agent will help an application hold one of its user accountable but also the fashion in which it will do so makes identity escrow a particularly promising option for users and applications looking to customize accountability mechanisms to their particular preferences and concerns. As discussed in the previous chapter, customization of this sort can be valuable in helping design accountability mechanisms suited to the great diversity of available Internet applications as well as the great diversity of user priorities.

8.4.2 Scoped Identities

Inserting a trusted third party into the relationship between end-users and applications also allows for the possibility that escrow agents storing users' full identity profiles might be able to reveal only certain, relevant identity elements to applications. In this scenario, users would again provide more thorough identity credentials to a trusted third party which would then scope that identity, or tailor it, to the specific needs of a given application. For instance, an application might only need to know that its users were at least 18 years old, and no additional information about their identities, or even their specific ages. In this case, a third party could simply indicate to the application whether or not given users were 18, providing no further identifying information beyond that scope. Similarly, some applications might want to know the nationality of each of its users, in order to tailor their content to different jurisdictional regulations in a more rigorous manner than can currently be achieved using IP address tracing techniques. Trusted third parties could again be called on to release only the citizenship of users to that application, without divulging any additional details.

These scoped identities, in which third parties reveal to applications only that users belong to specific groups—the group of people older than 18, or the group of people who are U.S. citizens—rather than their individual identities can afford users a fairly high degree of anonymity. Lessig (2006, p. 51) notes that scoped identities have the potential in some cases to provide even greater privacy protections for users' identities than they enjoy in the physical world, writing:

> In real space, if you want to authenticate that youre over 21 and therefore

> can buy a six-pack of beer, you show the clerk your drivers license. With that, he authenticates your age. But with that bit of data, he also gets access to your name, your address, and in some states, your social security number. Those other bits of data are not necessary for him to know ... The virtual wallet would be different. If you need to authenticate your age, the technology could authenticate that fact alone—indeed, it could authenticate simply that youre over 21, or over 65, or under 18, without revealing anything more ... The technology is crafted to reveal just what you want it to reveal, without also revealing other stuff.

Combining these privacy-protecting scoped identities with the conditional anonymity mechanisms of identity escrow agents can be a particularly effective means of creating anonymous-accountable online identities. An application relying on such a system can satisfy itself that it knows enough about its users' identities to allow them anonymous access to its services, while still being reassured that, should it need to enforce some more stringent accountability mechanism, it can turn to the escrow agent to revoke users' anonymity. It is important to note, though, that in some cases, the supposedly anonymous attributes revealed under scoped identity schemes may, in fact, provide sufficient information to allow for de-anonymization of users (Narayanan & Shmatikov, 2008).

8.4.3 Linking Identities

Identity escrow agents may extend their services to numerous different applications, in which case users' escrowed identities may be linked to multiple, separate anonymous identities associated with the different applications they use. This, in turn, creates the potential for linking together—through the escrow agent—the otherwise discrete online identities belonging to an individual user. In some instances this may be a valuable tool for pursuing criminal investigations or extending the reach of accountability mechanisms to span multiple applications, however, it can also present a risk to users' privacy and the customization of appropriate application-specific accountability mechanisms. The accountability ramifications of linking together the different application-specific online identities associated with an individual user are discussed in greater detail in the following chapter.

8.4.4 Escrow Agents & Specialization

Implemented well, identity escrow mechanisms can help reassure users that their real identities are being protected while alleviating applications of the burdens of collecting and storing identifying information for their users. In doing so, identity escrow creates an additional point of control for Internet applications—the escrow agent. Just as applications may struggle with authenticating users' identities, storing those identities securely, and making case-by-case judgments about when and how to use that information, so, too, may escrow agents. The primary advantage of transferring these burdens to a dedicated escrow entity instead of an individual application lies in the fact that the escrow agent can specialize in areas like identity authentication and data security while applications are often focused on other functions and therefore unable to devote as much time or effort to protecting users' identities. Shifting these responsibilities to a trusted third party can allow applications to focus more on developing their specific services. Still, despite these benefits, the addition of a new control point can complicate the trust relationships between applications and their end-users by introducing another potential source of vulnerability and insecurity.

8.5 Enforcement Challenges & Jurisdiction

Conditional anonymity schemes can provide a useful alternative or supplement to the accountability mechanisms offered by the investment-privilege identity trade-offs described in the previous chapter. In certain cases, providing a real identity—either directly to an application or to a trusted third party—can be viewed as a sort of investment on the part of the user since it allows the application operator to trace misbehaving identities and hold their real-world counterparts accountable. Using conditional anonymity as a means of holding users accountable can be made more difficult, however, by issues of legal inconsistencies and jurisdiction. Just because an application, or an escrow agent for that matter, can successfully identify the person responsible for some malicious activity within an online application does not mean that it will necessarily be easy or feasible for the application owner to hold that person accountable. If the offending party is located in a different jurisdiction than the application, and particularly if the user's behavior is not viewed as criminal or inappropriate within their native jurisdiction, it can be extremely difficult for an application to rely on legal or judicial recourse as a means for holding users accountable.

These challenges are inherent in attempts to use conditional anonymity schemes for the purposes of external accountability. However, there are also ways of using conditional anonymity schemes to augment internal accountability mechanisms. Charging users fees for misbehavior or using their real identities to prevent them from continually creating new anonymous online identities can sometimes suffice to enforce an application's accountability needs without involving any legal or jurisdictional element. From the perspective of application design, however, it is important to bear in mind that simply creating a mechanism for obtaining a user's real identity is not, in itself, a guarantee that it will be possible to hold that user accountable. Accountability is distinct from attribution and careful consideration and analysis of design decisions is needed to determine the most appropriate and effective method of leveraging a user's real identity to hold them accountable for past misbehavior and prevent future infractions.

Chapter 9

Aggregation, Federation & the Identity Layer

> As users, we need to see our various identities as part of an integrated world which none the less respects our need for independent contexts.
>
> —Kim Cameron, "The Laws of Identity" (2005)

One of the primary, underlying arguments of this analysis has been that different Internet applications require different approaches to identity, with accountability and anonymity mechanisms tailored to their specific functions and design goals. With the rapid proliferation of online applications, however, maintaining different, separate accounts, or identities, for each one can be both inconvenient and insecure for many users. Additionally, it can place a tremendous burden on application designers and operators to collect and protect identity information in a way that affords users an acceptable level of privacy while still enabling a reasonable degree of accountability. In the previous chapter, we discussed identity escrow mechanisms as one possible means of partially relieving applications of this burden by allowing them to rely on specialized online identity providers. Escrow agents are just one possible way to create online identities that can span several different applications. Other mechanisms such as website single sign-on accounts, centralized Internet identity providers, and federated identity management systems also provide users with a way to access multiple independent applications by means of a single online identity that exists, in some sense, just below the application layer—at the so-called "identity layer."

At first glance, the notion of an identity layer may appear at odds with the argument that different applications require different, customized approaches to online identity and accountability. However, identity management systems can, in fact, allow for such customization and careful tailoring of appropriate identity schemes and accountability mechanisms, while still relieving designers of some of the responsibility for creating these schemes and mechanisms from scratch for every new Internet application. For instance, identity layer schemes can be used to enable scoped identities, bring users' behavior in multiple different applications to the attention of their social networks, and expand the reach of individual application accountability mechanisms to impact numerous different virtual identities belonging to a single user. Thus, specialized identity management services that provide identity authentication and credentials to other online applications can be a valuable resource for application design. This chapter looks at several different identity management models, from application-specific identity silos to centralized and federated identity layer mechanisms, and analyzes the ramifications of each model for application-layer anonymity and accountability.

9.1 Laws of Identity & Accountability

Before evaluating different models of Internet identities, it is helpful first to consider what general traits an online identity system should and should not exhibit to promote accountability. Cameron (2005) proposes seven overarching "laws of identity" for digital identity systems. These include:

1. User control and consent: users must agree to any disclosure of their information by an identity system.
2. Minimal disclosure for a constrained use: An identity system should always disclose the minimum amount of identifying information that is needed for a user to get access to another application.
3. Justifiable parties: An identity system should only disclose identifying information to parties that require access to it for a justifiable reason.
4. Directed identity: An identity system should allow for both publicly broadcast identifiers and privately concealed identifiers.
5. Pluralism of operators and technologies: An identity system should allow for the coexistence of and interoperation between multiple identity technologies and providers.
6. Human integration: An identity system must incorporate consideration of its human users and ensure that they can reliably and easily use and communicate with the system.
7. Consistent experience across contexts: An identity system should provide users with a consistent experience while still allowing for separation of different contexts in which different identity credentials may be needed.

Cameron's seven laws provide a useful starting point for analyzing online identity models, particularly if we refine them to focus more specifically on the desirable traits of accountable identity systems. For instance, Cameron's notion of user control and consent is related to the idea that users should be aware of the accountability mechanisms used by a given application and informed of under exactly what circumstances any identifying information they provide may or may not be used to hold them responsible for their actions. For obvious reasons, it is probably not effective or desirable to ask a user to consent to a disclosure of their information after they have engaged in some form of malicious online activity. However, it is important for application designers to outline as specifically as possible to users what types of application activity will result in such disclosures or other identity-based disciplinary measures. Similarly, Cameron's law of minimal disclosure ties into the idea that an identity system should collect and disclose only as much identifying information as is needed for a user to be held accountable for their actions. Often, as discussed in the previous chapter, this may involve revealing minimal amounts of information about well behaved users but more information about misbehaving users, depending on the seriousness of their infractions. In this manner, we can develop seven laws of accountable identity systems which are loosely related to Cameron's ideas but more specifically focused on designing for accountability.

1. User awareness and consent: An identity system should clearly inform users of the specific circumstances under which elements of their identifying information may be used, and in what way, to punish malicious activity and prevent further infractions. In particular, users should be made aware of the ways in which their identifying information will be scoped to suit the needs of different applications and be given control over how their actions within one application may be publicized or recorded by another.

2. Proportionality of disclosure and misbehavior: An identity system should disclose minimal identifying information about users who conform to appropriate uses and behavioral norms for an application. In the event of misbehavior, an identity system may disclose more identifying information about the offending users, only as needed to hold them accountable for their actions and in proportion with the seriousness of their infractions.

3. Justifiable parties: An identity system should only disclose identifying information about misbehaving users to parties that are directly involved in the mechanisms designed to hold those users accountable for their actions. Depending on the accountability mechanism, these parties may include actors such as application operators, escrow agents, identity management systems, courts of law, and even the population of other users within an application.

4. Public & private accountability mechanisms: An identity system should allow for both accountability mechanisms that function as a public signal to other users about the reputation of one of their fellow users, as well as those that provide private indicators to an application operator or law enforcement agency about the real identities of malicious users. Similarly, an identity system should provide mechanisms for both internal and external accountability.

5. Pluralism of accountability tools: An identity system should allow for the coexistence of multiple different accountability mechanisms and, where possible, provide users with choice about which of those tools they wish to be subject to, and to what extent.

6. Human integration: An identity system must incorporate consideration of whether its accountability mechanisms will appropriately impact its human users, how clearly these mechanisms will be communicated to users, and how effectively they will serve as deterrents of misbehavior.

7. Consistent experience across different accountability contexts: An identity system should provide users with a consistent experience while still allowing for a user who is held accountable for misbehavior in one context, or application, not necessarily to be held accountable for that same misbehavior in other contexts or applications, except where appropriate. In other words, malicious activity within one application should not automatically trigger accountability mechanisms for that same user in other applications, even if that user is using a common identity management system for both applications. However, in many cases it may be valuable to offer users the option of electing to be simultaneously held accountable within the separate contexts of multiple applications.

This seventh principle warrants some further discussion to clarify the ways in which linking online identities via identity management and aggregation systems can enable stronger forms of accountability.

9.1.1 Linking Identities for Accountability

Linking identity investment across different applications may provide very promising possibilities for implementing strong accountability mechanisms. For instance, a key question is whether malicious activity in one application should affect a user's reputation (or privileges) in other applications that also use that same identity. Such schemes can be criticized for invading users' privacy by making the consequences of their online misdeeds too far-reaching, beyond the specific context in which they were perpetrated. However, these systems also have the potential to be very powerful accountability mechanisms since they could greatly amplify the consequences of online misbehavior and therefore act as more effective deterrents. In some sense, allowing a single online identity to be linked across multiple applications can be viewed as a further form of investment in that identity. In these instances, users are not only willing to invest time and energy in building up an identity for a specific application, they are also willing to stake their reputations and identity status in several other applications on that same, single identity. This could serve as a fairly strong signal to application operators of how much a user has invested in a given identity and therefore how accountable that identity is, and how many privileges it should be given, accordingly.

Therefore, one possible solution to the concerns about the privacy implications of linking online identities across applications is to give users the choice of whether they want their identity provider to share reputational information associated with their identities between applications. In this manner, users who valued privacy and detached identities most highly would not be forced to aggregate their online identities, while other users would be able to reap the benefits of their previously established online reputations without having to go to the effort of building up an entirely new online identity. Users who chose not to export their previous online identities and associated investments might be viewed with greater suspicion in new applications and be afforded fewer privileges, but this would be a natural and logical extension of the previously described investment-privilege trade-off framework.

Inter-application identity sharing systems could have tremendous accountability ramifications, in addition to the considerable user convenience and other benefits they offer to users and application operators. Giving users the option to link online identities across multiple different applications creates the possibility of implementing accountability mechanisms in which loss of privileges or status in one application reverberates to include loss of privileges and status in the other applications using that same identity. This enables application designers to leverage the accountability mechanisms of other applications to amplify the effects of their own tools for holding users responsible for misbehavior. It also affords end-users the luxury of not having to invest over and over again in new identities for each new application they join in order to gain full privileges from every online service they use. Instead, users would have the option of taking advantage of previous identity investments to enjoy greater privileges in new applications. Making this sharing of identities a choice, rather than requiring it, would allow users to continue to maintain some diversity of online identities, including the potential for siloed application-specific identities that they wish to keep separate from the rest of their other online identities. Even though some applications and some end-users may require separate or unlinked identities for various reasons, many others may be able to integrate their tailored accountability mechanisms, especially by tying application-specific privileges to the identity investments made by users in other applications and online contexts. This has the dual benefit of making investment in online identities more meaningful and worthwhile for end-users and also rendering the application-imposed consequences of malicious activity more significant and far-reaching.

9.2 Application Identity Silos

The seven accountability principles laid out in the previous section inform our further analysis of different models of identity management. Notably, these principles allow for the preservation of the central tenet of "application-layer accountability"—diversity of online identities and accountability mechanisms appropriate to the diversity of Internet applications—even at the slightly lower identity layer. To understand how this works, it is worth reviewing the "silo" approach to application identities that underlies most of the application-layer mechanisms discussed thus far.

The original model of online identities for Internet applications involves each application providing its own set of users with an application-specific identity. These identity "silos" allowed every online service to implement individual authorization and identification requirements, including the corresponding accountability mechanisms (Priem, Leenes, Kosta, & Kuczerawy, 2011). From the perspective of application designers, this model has the advantage of allowing for customization of appropriate, carefully tailored identity and accountability schemes for the needs and purposes of each specific application. Furthermore, it enables each application to collect potentially useful or valuable data and identifying information about their users. However, it also has the disadvantage of requiring each individual application designer to create and implement an appropriate and effective identity mechanism, creating an entirely new dimension to the design challenges they face, and possibly one outside their interests or area of expertise. From the perspective of end-users, there are also clear advantages and disadvantages to the silo model. One positive consequence of maintaining a different identity for each different application is that end-users can customize their own identities for different applications, choosing how much of their real identity they wish to reveal as well as what specific traits they want to espouse for each online application they use. More negatively, however, as the number of available applications grows and users create identities for more and more of them, maintaining all these different accounts and remembering the necessary authentication information for each one can become burdensome to end-users. Additionally, creating accounts for new identities often requires end-users to enter the same personal information repeatedly (e.g., name, address, credit card number, etc.) which may be both tedious for the users as well as potentially insecure.

This set-up, in which applications, or service providers, also act as identity providers has several desirable accountability attributes. Since users input identifying information for each specific application, they know exactly what information each service provider has access to and can use for accountability purposes. Terms of Service agreements and privacy policies may still play an important role in informing users of exactly how and when that information may be used, but end-users at least have a clear sense of what identifying information they have provided within each "silo," or application. Additionally, users are assured of having separate identity and accountability experiences in different application contexts since misbehavior, and the associated consequences, within one application will have no bearing on the identities held by that same user in other applications. For application designers and operators, however, there can be downsides to this strict separation of contexts: It may be valuable for a given application to know about the ways in which its users have misbehaved and been held accountable by other applications. In general, though, improving the experience of end-users, rather than application designers, has been the strongest driver of the implementation of aggregated identity-layer mechanisms that span multiple applications.

A 2007 study showed that Internet users maintained roughly 25 different password-protected online accounts, entered an average of 8 passwords per day, and regularly forgot these passwords and were forced to reset them (Florencio & Herley, 2007). As end-users accumulate more and more online identities, they may experience "password fatigue" and become overwhelmed by how many different passwords and accounts they must remember and maintain under the silo model (Sun, Boshmaf, Hawkey, & Beznosov, 2010). As the isolated identity silo model became unmanageable for many users with the increase in available online applications, some companies began experimenting with centralized single sign-on (SSO) identities that would allow users to access multiple different applications using the same, single identity. These experiments with SSO accounts spawned an extensive body of research on identity management system design, ranging from centralized and federated models to mechanisms that afford end-users differing degrees of control over their identity attributes. While there is considerable literature on the technical implementation of such schemes, as well as their privacy and security ramifications, relatively little work has been done focusing on their effects on online accountability. In the following sections we explore some of the different models of identity management and aggregation, including centralized SSO identities and federated models, paying particular attention to the types of accountability afforded by these mechanisms.

9.3 Centralized Single Sign-On Identities

One of the earliest forms of identity management was for companies and networks that maintained multiple different applications to allow their users to access all of their services through a single, centralized account. The Kerberos authentication protocol for services provided on the same network follows this model, issuing users tickets that authenticate them to all the different network services operated under the same administrative control (Kormann & Rubin, 2000). Online, a company that offers an e-mail service, a chat service, and a social networking service, could take a similar approach by allowing a user to access all of these different applications using the same account. This form of identity management, called single organization single sign-on (SOSSO), may benefit users by mildly reducing the burdens associated with remembering information for multiple different accounts and benefit the application designers and operators by providing the opportunity to link individual users' actions across multiple different applications (Priem et al., 2011). This has potential to be a useful trait for accountability mechanisms, since it suggests that a user who misbehaves within one application could also be flagged or punished across several other applications where one might worry they would also cause trouble. However, with the current diversity of companies owning and operating online applications, SOSSO solutions are very limited both in how much convenience they offer users and how much potential for increased accountability they provide to application designers. Furthermore, the SOSSO model of identity management still requires each new company to design and implement its own identity system, so it does not noticeably alleviate the accountability design burdens placed on many designers looking to create new applications.

From SOSSO technology, a natural next step was multi-organization SSO systems, like Microsoft's Passport program. Passport allowed users to access many different applications owned and operated by a variety of different companies using a single Passport account. Users' identity information is all stored centrally, by Microsoft, but once these users were authenticated by Microsoft they were also automatically considered authenticated by other service providers who joined the program and agreed to accept the Passport identity credentials (Jøsang & Pope, 2005)). In other words, end-users trust Microsoft to collect and maintain their identifying information and other applications, in turn, trust Microsoft to authenticate their users for them. When Passport users try to login to one of the program's partner sites, they are redirected to Microsoft's login page where they are given an authentication cookie that can be used to access their desired application, as shown in Figure 9-1. Each partner site can designate how much identifying information users must provide to Microsoft in order to access it (the minimum amount being an e-mail address and password) and each user can, in turn, designate which pieces—if any—of this profile information Microsoft may share with the partner site by means of an additional "profile cookie." Thus, both partner applications and end-users are able to customize their desired levels of authentication information and privacy, respectively, enabling the creation of application- and user-specific identity and accountability mechanisms within the broader, overarching Passport system. At the same time, the Passport system, by design, requires all partner sites to go through Microsoft to authenticate users, so the Passport model intrinsically violates Camerons law of justifiable parties since, even when there is no justifiable reason for Microsoft to be involved in these transactions and know which applications its users are trying to access, there is no way of avoiding disclosing that information to them (Chadwick, 2009).

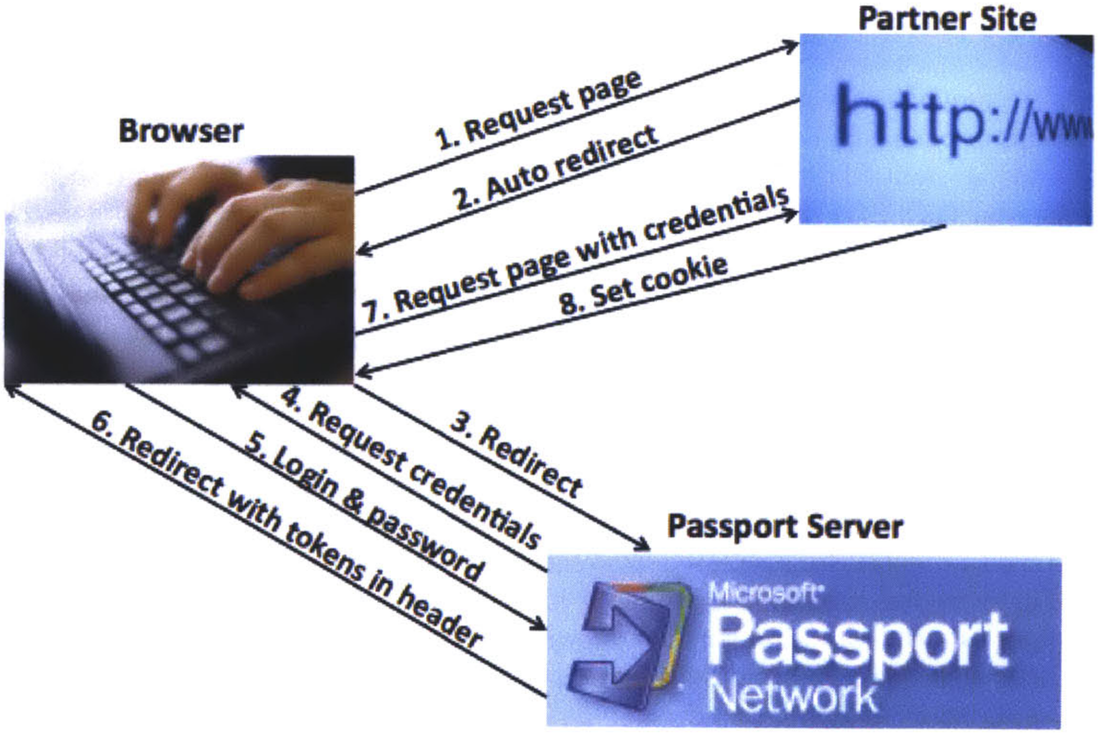


Figure 9-1: The architecture of Microsoft's Passport program.

Multiple-organization SSO models have some clear advantages over the single-organization SSO and identity silo set-ups. First, if a large number of applications are willing to trust a single other party to perform authentication, they can dramatically cut down the number of online identities a single end-user needs to maintain in order to access a variety of applications. Additionally, multi-organization SSO enables application designers to outsource their authentication and identity mechanisms to a trusted party which may be more specialized or experienced in the area of identity management. The convenience these schemes offer to end-users and application designers alike, however, is tempered by the reluctance of some application designer to relinquish their users' profile information to a third party. Thus, from an accountability standpoint, these centralized multiple-organization SSO systems can enable the development of application-specific accountability mechanisms by allowing applications to determine what authentication information must be collected about its users while letting users decide whether or not to share that information directly with the application. This can allow for the implementation of scoped identities, in which users agree to release certain identity attributes to specific applications, without revealing their real identities. Since users' different virtual identities are stored centrally by a single entity, it may also be possible for that entity to link these identities so that changes to the status of a user's identity made for one application apply to all the other applications using that same identity for authentication purposes. This can be a valuable trait for creating more powerful accountability mechanisms that resonate across multiple applications, though it may require these applications to implement similar—or at least interoperable—accountability schemes in order for identity status changes in one application to be reasonably translated to similarly impact identities in other applications.

Although they exhibit some promising potential for accountability mechanisms, multi-organization SSO systems have come under fire for posing some serious security risks. SSO systems, which would ideally be designed to specialize in securing identity information and operated by experts in this area, often have major security flaws that allow attackers to sign in under other users' identities. One recent study identified 8 such flaws in a range of high-profile SSO providers, including Facebook, Google ID, PayPal Access, Freelancer, JanRain, Sears and FarmVille (Wang, Chen, & Wan, 2012). Several researchers have pointed out that centralized SSO schemes are particularly vulnerable because they feature a single point of failure—the login server—which is likely to be the target of denial-of-service attacks as well as attempts at unauthorized access (Kormann & Rubin, 2000). Thus, the security protections afforded by reducing the number of applications that store user identity and authentication information may be mitigated, or even outweighed, by the dangers of centrally storing all of that data in only one place. Decentralizing identity management systems, so that users can choose among many different identity providers while still enjoying the convenience of SSO, is one means of trying to allay concerns about these single-point-of-failure vulnerabilities. This decentralized, or federated, identity management model offers users greater choice of identity providers, but offers fewer benefits to applications than many centralized identity systems.

9.4 Federated Identity Models

Like centralized identity models, federated identity management systems can provide users with SSO ability that reduces the burdens of maintaining numerous different identities and remembering the associated authentication information. However, unlike centralized models, in federated identity domains different applications all agree to recognize users who have been authenticated by any other partner service in the federation. In other words, when a user's identity is authenticated by one application in the federation, the rest of the applications in that federation will automatically accept that same identity as authenticated (Jøsang & Pope, 2005). Systems like Shibboleth and OpenID follow this decentralized model, in which users can select their own identity provider, instead of being forced to using a single centralized provider. Since there is no central repository for all users' identity information, the federated model does not feature the same single-point-of-failure vulnerabilities as the centralized model, however, these identity management schemes can still pose grave security risks. Relying parties in a federation typically redirect a user's browser to their chosen identity provider for authentication and that identity provider must then pass the user's browser a secure token that can be used to authenticate their identity to the target application. If the process of passing that token from the identity provider to the browser and then to the relying party is not sufficiently secure, it can allow attackers to gain unauthorized access to other users' identities (Wang et al., 2012).

The differences between centralized and federated models in terms of user accountability are fairly minimal, though federated identity systems may complicate the issue of accountability slightly since there is no single authority responsible for maintaining users' identities and holding them accountable for malicious activity. As in the case of the centralized model, it is possible for an application to try to hold a misbehaving user accountable by going to that user's identity provider and requesting assistance in identifying and punishing the user, but in federated systems, applications must maintain multiple relationships with a variety of different identity providers. To help clarify these relationships, federated systems like Shibboleth allow both users and identity providers to set "Attribute Release Policies" governing which identity attributes may or may not be released to relying applications, allowing for tailored and scoped identities in a manner similar to Microsoft's Passport system (Chadwick, 2009). Federated and centralized identity providers also have in common the capability to link different virtual identities associated with the same user, though in federated systems, where more identity providers are available for users to choose between, it may be easier for end-users to maintain multiple different, unlinked identities by entrusting each one to a different identity provider. Federated identity systems, like centralized models, also pose drawbacks to application providers who do not wish to give up access to their users' identifying information.

Adoption of SSO systems—whether centralized or federated—has been hindered by several concerns and competing technologies. For instance, one study of users found that more than one-fifth of participants used their browser's password management service to remember authentication credentials for their different online identities, essentially turning the browser into a simple identity management tool (Sun et al., 2011). Participants also expressed concerns about single-point-of-failure security issues and phishing attacks in which they would be redirected to fake login pages that would capture their authentication information. Notably, half of the respondents were unable to distinguish a fake Google login page from an authentic one. Without strong support and widespread adoption from end-users, it is unlikely that applications themselves will drive a move towards more aggregated identity management systems, even though such mechanisms could potentially relieve them of some of the burden of collecting and protecting their users' identities. Bonneau and Preibusch (2010) note, "deployment of an open, federated identity protocol such as OpenID will be opposed by current stakeholders on the web. Federated login not only removes sites' pretext for collecting personal data but their ability to establish a trusted relationship with users."

9.5 The Future of Online Identity Management

Though many end-users are content relying on password managers instead of more sophisticated identity management tools, and many applications are unwilling to lose the personal data they collect about their users, there is one model of identity management that seems to hold some promise for the future: popular social media applications acting as centralized identity providers. These multi-organization SSO systems, like Facebook Connect and Google ID, allow users to log in to other applications with their Facebook or Google accounts and can simultaneously enrich users' application experiences and enhance application owners' ability to collect in-depth data about their user base. Indeed, there is some evidence that these mechanisms are growing more popular as end-users and applications alike reap the benefits.

Since its launch in 2008, more than one million different relying parties had begun allowing users to log in to their applications via Facebook Connect and more than 250 million people used the service every month (Sun et al., 2010). By connecting to other applications through Facebook, end-users can connect with their Facebook friends in these other applications by offering or receiving recommendations and posting achievements and activities. Facebook, in turn, can collect even more data about what services and other applications are popular with its users. Finally, relying parties can take advantage of Facebook-provided user information and also the marketing potential of social media communities, by directing Facebook advertisements and targeting Facebook friends of current users.

Clearly, there are some circumstances and some applications for which such a social media-oriented identity system would not be desirable. In particular, social media identities rarely afford a high degree of anonymity since they are focused primarily on connecting people with their real-life friends and social circles. However, for applications where users do not desire strong anonymity and actively enjoy being able to interact with the social circles they've defined in other applications, this model can be a very advantageous way to minimize the number of different online identities an individual user must maintain. These shared inter-application identities may even be pseudonymous since the key factor is not the degree to which an online identity is linked to its user's real identity but instead the degree to which the user has invested in that online identity.

Facebook Connect is much more than an authentication mechanism. It provides benefits to end-users and applications alike that go well beyond simply protecting identity information and ensuring that users are who they claim to be. Where systems like Microsoft Passport and Shibboleth are intended solely to authenticate users to different applications with greater convenience, Facebook Connect serves as a much richer identity tool in which users can transport their identity information, including social connections, reputation, and more, across a variety of different applications. Authentication is, in some sense, almost secondary to the other benefits this can provide to users and applications. Returning to our earlier discussion of investment in online identities, it seems unsurprising that users might want to be able to transport these profiles as they are investing more and more time in building them. Password fatigue aside, it can be tedious and time-consuming for end-users to have to invest substantial time (or money) in building up an identity for each new application they use. These investments are crucial to rendering online identities less discardable and therefore more accountable, but they require considerable time and effort on the part of end-users. The earlier case studies illustrate how some applications have successfully encouraged a large user base to invest in certain application-specific identities, from Facebook profiles to Yelp reputations. The move towards identity management systems like Facebook Connect may hint at those applications realizing just how valuable that investment is, not just to them but potentially to other applications as well.

Online identities that users have invested heavily in are valuable commodities to Internet applications. By allowing users to login through established applications like Facebook, designers of new applications can tailor the privilege side of the previously described investment-privilege trade-offs to their own applications, while still taking advantage of the investments users have already made in identities established with other applications.

Chapter 10

Conclusion

> One way or another, people always wind up being held accountable.
>
> —U.S. President Bill Clinton
> June 13, 1996

Efforts to embed greater accountability in the Internet have ranged from designing protocols that prevent IP spoofing (Mirkovic & Reiher, 2008; Andersen et al., 2008) to calling for network-layer attribution mechanisms that enable tracing all packets back to specific users (McConnell, 2010; Davenport, 2002). Implementing these network-layer approaches pose some serious drawbacks: even when it is possible to reliably trace packets back to their originating machines, it may be impossible—or difficult—to identify the responsible user, and even if that user can also be successfully traced, jurisdictional borders may hinder that attribution process, as well as subsequent attempts to hold the identified user accountable. Furthermore, an accountability measure that is applied uniformly to the entire Internet cannot provide a variety of different types and degrees of accountability mechanisms suited to the current diversity of online applications. Application-layer approaches to accountability, by contrast, allow for tailoring specific accountable-anonymous identity mechanisms to the functionality and architecture of different Internet applications.

10.1 Anonymous-Accountable Online Identities

Several relevant observations emerge from a close analysis of these application-specific accountability mechanisms. These conclusions include:

- Anonymity and accountability are not mutually exclusive online. It is possible to design accountable-anonymous Internet identity schemes that incorporate elements of both anonymity protections and some accountability mechanisms.
  - Implementing accountability mechanisms at the application layer allows for the development of these accountable-anonymous schemes in a variety of ways tailored to suit different types of applications and the types of accountability and anonymity best suited to their form and functions.

- Effective accountability mechanisms require forcing users to invest in their online identities to reduce the discardability of those identities. Users who are more heavily invested in their virtual identities can be more effectively held accountable for their online actions.

- Identity investment-privilege trade-offs can allow applications to relax behavioral constraints placed on users who are more invested in their online identities, allowing for internal accountability mechanisms that take advantage of the diverse types of investments and privileges enabled by different applications.

- Conditional anonymity schemes can enable both internal and external accountability by revoking anonymity protections of only those users who engage in malicious activity.

- Identity-layer systems that manage or aggregate online identities can allow for inter-application accountability schemes in which a user's misbehavior in one application has ramifications for the status of that same user's other online identities, as well.

    - Users who allow their different identities across multiple applications to be linked together for accountability purposes can be viewed as making a greater investment in their virtual identities and, accordingly, afforded greater privileges.

- It is often advantageous to allow for greater customization of accountability schemes both by different applications and by their individual users. In this manner, application designers can define an acceptable array of accountability options tailored to their applications and individual end-users are then able to select which of those options is best suited to their preferences. Such mechanisms can maximize both designer and end-user choice and help meet the diverse accountability and anonymity needs of the wide variety of available online applications as well as the billions of Internet users worldwide.

10.2 Layers of Accountability

There are many different types of accountability on the Internet, as in the physical world, including vertical and horizontal accountability, as well as internal and external. In several of the different accountability mechanisms discussed thus far, these different kinds of accountability are combined in ways intended to reinforce each other and strengthen the overall ability of applications to hold users responsible for their actions. A related approach to combining different types and degrees of accountability online involves embedding different layers of accountability within online identities.

Farkas et al. (2002) propose a two-layer system for accountability of online identities, in which the first layer associates a user's real identity with a single "base" virtual identity and the second layer associates that user's various different online identities with that base identity. Both associations are encrypted and can be revealed only when deemed necessary by both a trusted computing base and a group of trusted users. The second layer of accountability thus enables applications to determine whether two different virtual entities are controlled by the same user, while the first layer allows for tracing the real identity of associated with a virtual identity, when necessary. These different layers can then be leveraged depending on how serious a user's misconduct is and whether it calls for holding just the user's virtual identity accountable, by banning him from creating additional identities within an application for instance, or instead warrants holding a real person accountable via legal action or other external mechanisms.

Designing online identities that incorporate different layers, or degrees, of accountability in this manner has the potential to be a useful tool for ensuring that even within individual applications, it is possible to adjust the ways in which malicious actors are held accountable depending on the severity of their actions. This allows for the possibility of tailoring accountability mechanisms not only to the functions of different applications and the preferences of individual end-users, but also to the type of misconduct for which a user is being held accountable.

10.3 The Way Forward For Online Accountability & Anonymity

A growing number of governments and individuals like McConnell (2010) argue that weaker accountability mechanisms are too high a price to pay for the anonymity afforded by the Internet. Davenport (2002, p. 35) goes so far as to declare: "The way forward is clear: embrace accountability and reject anonymous communications." But calls like these for network-wide attribution capabilities have been countered by increasing pressures to improve and strengthen anonymity protections, demonstrating that the way forward is far from clear for online identities. Even within individual national governments, there is often disagreement about the best course of action for the future of anonymity and accountability on the Internet. While McConnell, a former officer of the U.S. Department of Defense, advocates for affixing license plates to packets, the U.S. Department of State has provided millions of dollars in funding for anonymity-protecting technologies that help dissidents and political activists living under oppressive regimes communicate securely and avoid censorship (Glanz & Markoff, 2011).

This ongoing tussle and the lack of any clear resolution between the two sides highlights the need for further analysis of how accountability and anonymity can best be reconciled to preserve the rich diversity of Internet applications available today and allow for even more online innovation to emerge in the future. Levmore (2010, p. 67) predicts:

> Over time ... more Internet entrepreneurs will limit participation or require identification. There will remain a few sites ... occupying a niche like that in which we find porn shops in city centers, at the periphery

> of most social interaction . . . 'Respectable' sites will require identification (non-anonymity) and this will severely limit sites where people comment on a professor or classmates anatomy or alleged promiscuity. There will be some loss of opportunities to flatter, criticize, and convey information. But inasmuch as this information would have been lost in the midst of much noise, most of us will not and should not mourn the loss.

However, even with the growing prevalence of social network identities being used by other applications through programs like Facebook Connect and Google ID, there is evidence that a significant number of Internet users would, in fact, mourn the total loss of their online anonymity.

Facebook and Google+ pride themselves on being able to identify the real names and identities of the large majority of their users though, as discussed in chapter 5, it is not clear that their accountability mechanisms are primarily dependent on being able to link their users to real identities. It remains possible—though sometimes difficult—for users of these social networking applications to maintain their anonymity to some degree, but for many Internet users, particularly those in cultures that highly value online anonymity, these opportunities may not be adequate. In Japan, for instance, Facebook has struggled to gain a prominent presence among other, competing social network sites which explicitly permit pseudonymous accounts. In 2011, Facebook had fewer than 2 million Japanese users (or less than 2 percent of the country's Internet users), while popular Japanese social network applications Mixi, Gree, and Mobagetown—none of which require users to provide their real names—each boasted more than 20 million. Even Twitter, the American micro-blogging application which does not require linking handles to users' real identities, has attracted about 10 million users in Japan, and cell phone surveys indicate that roughly 89 percent of Japanese Internet users are "reluctant to disclose their real names on the Web" (Tabuchi, 2011).

Real name policies have met with resistance in other countries as well. In Saudi Arabia, women often create social network profiles under fake names following an incident in 2008 when a Saudi Arabian woman was killed by her father for using her Facebook account to chat with a man. Similarly, in 2009 several Iranian users changed their last names on their Facebook profiles to "Irani" in the wake of Iranian government officials harassing citizens for their social media activities (Harris & Llanso, 2011). As Schneider (2009) points out:

> Various cultures resolve tension between anonymity and accountability in different ways, perhaps even selecting different trade-offs for their own traffic than for outsiders traffic. In short, theres no universal agreement on mandates for accountability.

For social network-based identity services like Facebook Connect and Google ID to achieve widespread dominance it may therefore be necessary for them to make greater allowances for user anonymity. Their current policies of strongly—though not always successfully—discouraging pseudonyms appear unlikely to provide the basis for a comprehensive, global Internet identity ecosystem. Such an ecosystem may incorporate several of the different techniques discussed previously, including identity investment-privilege trade-offs, conditional anonymity schemes, scoped identities, and aggregated identity management systems. No individual method or form of identity is likely to suffice for constructing this ecosystem; it will require designing mechanisms to enable the appropriate combinations and layers of accountability and anonymity for virtual identities that meet the needs and preferences of different cultures, users, applications, and types of online misconduct, from theft to trolling.

References

Afanasyev, M., Kohno, T., Ma, J., Murphy, N., Savage, S., Snoeren, A. C., et al. (2011, May). Privacy-preserving network forensics. *Communications of the ACM*, *54*, 78–87. Available from http://doi.acm.org/10.1145/1941487.1941508

Aiken, C. T. (2008, June). Sources of Law and Modes of Governance—Ethnography and Theory in Second Life. *University of Pittsburgh Journal of Technology Law & Policy*, *10*. Available from http://tlp.law.pitt.edu/ojs/index.php/tlp/article/view/55/55

Andersen, D. G., Balakrishnan, H., Feamster, N., Koponen, T., Moon, D., & Shenker, S. (2008, August). Accountable Internet Protocol. *SIGCOMM Comput. Commun. Rev.*, *38*(4), 339–350. Available from http://doi.acm.org/10.1145/1402946.1402997

Anderson, C. (2008, February). Free! Why $0.00 Is The Future of Business. *Wired*, *16*(3). Available from http://www.wired.com/techbiz/it/magazine/16-03/ff_free

Ansfield, J. (2009, September 6). China Web Sites Seeking Users' Names. *The New York Times*. Available from http://www.nytimes.com/2009/09/06/world/asia/06chinanet.html

Antin, J., & Churchill, E. F. (2011, May). Badges in Social Media: A Social Psychological Perspective. In *Proceedings of the 2011 SIGCHI Conference on Human Factors in Computing Systems.* Vancouver, BC, Canada.: CHI 2011, ACM. Available from http://gamification-research.org/wp-content/uploads/2011/04/03-Antin-Churchill.pdf

Applause Store Productions Ltd and Matthew Firsht v Grant Raphael. (2008, EWHC 1781 (QB)). Available from http://www.bailii.org/ew/cases/EWHC/QB/2008/1781.html

Atkins, L. (2011, Feb. 2). Goodmail shutting down. *Circle ID*. Available from http://www.circleid.com/posts/20110202_goodmail_shutting_down

Aura, T., & Ellison, C. (2000). Privacy and accountability in certificate systems. In *Res. Rep. A61.* Espoo, Finland: Helsinki University of Technology. Available from www.tcs.hut.fi/old/papers/aura/HUT-TCS-A61.ps

Bentley, A. (2011, June 3). Facebook troll set free. *The Brisbane Times*. Available from http://www.brisbanetimes.com.au/technology/technology-news/facebook-troll-set-free-20110603-1fjwb.html

Bethencourt, J., Shi, E., & Song, D. (2010). Signatures of reputation. In *Proceedings* *of the 14th international conference on financial cryptography and data security* (pp. 400–407). Berlin, Heidelberg: Springer-Verlag. Available from http://dx.doi.org/10.1007/978-3-642-14577-3_35

Bingham, J. (2008, July 24). Businessman awarded £22,000 damages over fake Facebook site. *The Telegraph.* Available from http://www.telegraph.co.uk/news/uknews/2453538/Businessman-awarded-22000-damages-over-fake-Facebook-site.html

Boellstorff, T. (2010). *Coming of Age in Second Life: An Anthropologist Explores the Virtually Human.* Princeton, NJ: Princeton University Press.

Bonneau, J., & Preibusch, S. (2010, June). The Password Thicket: Technical and Market Failures in Human Authentication on the Web. In *Proceedings of the 9^{th} Workshop on the Economics of Information Security.* Boston, MA: WEIS '10. Available from www.cl.cam.ac.uk/~jcb82/doc/BP10-WEIS-password_thicket.pdf

Bosker, B. (2011, July 27). Facebook's Randi Zuckerberg: Anonymity Online 'Has To Go Away'. *The Huffington Post.* Available from http://www.huffingtonpost.com/2011/07/27/randi-zuckerberg-anonymity-online_n_910892.html

Boyd, D. (2011, August). *Real Names Policies Are an Abuse of Power.* Available from zephoria.org/thoughts/archives/2011/08/04/real-names.html

Boyd, D., & Ellison, N. (2008). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication, 13*(1), 210-230. Available from http://jcmc.indiana.edu/vol13/issue1/boyd.ellison.html

Brennan, G., & Pettit, P. (2004). Esteem, Identifiability and the Internet. *Information technology and moral philosophy, 26*(1), 139-157.

Brownlee, J. (2007, March 2). John Edwards Meets Second Life 'Feces Spewing Obscenity'. *Wired.* Available from http://www.wired.com/table_of_malcontents/2007/03/john_edwards_me/

Cameron, K. (2005, May). The Laws of Identity. *Kim Cameron's Identity Weblog.* Available from http://www.identityblog.com/stories/2004/12/09/thelaws.html

Castronova, E. (2005). *Synthetic Worlds: The Business and Culture of Online Games.* Chicago, IL: University of Chicago Press.

Chadwick, D. W. (2009). Foundations of security analysis and design v. In A. Aldini, G. Barthe, & R. Gorrieri (Eds.), (pp. 96–120). Berlin, Heidelberg: Springer-Verlag. Available from http://dx.doi.org/10.1007/978-3-642-03829-7_3

Chaum, D. (1983). Blind Signatures for Untraceable Payments. In *Advances in Cryptology: Proceedings of CRYPTO '82* (Vol. 3, pp. 199–203). Santa Barbara, CA: Plenum Press. Available from ftp://zedz.net/pub/mirrors/Advances%20in%20Cryptology/HTML/PDF/C82/199.PDF

Chaum, D., Fiat, A., & Naor, M. (1990). Untraceable electronic cash. In *Proceedings on advances in cryptology* (pp. 319–327). New York, NY, USA: Springer-Verlag New York, Inc. Available from http://dl.acm.org/citation.cfm?id=88314.88969

Clark, D. D., & Blumenthal, M. (2011, March). The End-to-End Argument and Application Design: The Role of Trust. *Federal Communications Law Journal,* *63*(2), 357–390. Available from http://www.law.indiana.edu/fclj/pubs/v63/no2/Vol.63-2_2011-Mar._Art.-02_Clark.pdf

Clark, D. D., & Landau, S. (2010). The Problem Isn't Attribution: It's Multi-stage Attacks. In *Proceedings of the Re-Architecting the Internet Workshop* (pp. 11:1–11:6). New York, NY, USA: ACM. Available from http://doi.acm.org/10.1145/1921233.1921247

Clark, D. D., & Landau, S. (2011). Untangling attribution. *Harvard National Security Journal, vol. 2.* Available from http://harvardnsj.org/2011/03/untangling-attribution-2/

Clark, D. D., Wroclawski, J., Sollins, K. R., & Braden, R. (2005, June). Tussle in Cyberspace: Defining Tomorrow's Internet. *IEEE/ACM Trans. Netw.*, *13*, 462–475. Available from http://dx.doi.org/10.1109/TNET.2005.850224

Clinton, H. (2010, January). *Remarks on Internet Freedom.* The Newseum, Washington, DC. Available from http://www.state.gov/secretary/rm/2010/01/135519.htm

Curtis, P. (1997). Mudding: Social phenomena in text-based virtual realities. In S. Kiesler (Ed.), *Culture of the Internet* (p. 121-143). Mahwah, N.J.: Lawrence Erlbaum Associates, Inc. Available from http://w2.eff.org/Net_culture/MOO_MUD_IRC/curtis_mudding.article

Davenport, D. (2002, April). Anonymity on the Internet: Why the price may be too high. *Commun. ACM*, *45*(4), 33–35. Available from http://doi.acm.org/10.1145/505248.505267

Davida, G. I., Frankel, Y., Tsiounis, Y., & Yung, M. (1997). Anonymity control in e-cash systems. In *Proceedings of the first international conference on financial cryptography* (pp. 1–16). London, UK, UK: Springer-Verlag. Available from http://dl.acm.org/citation.cfm?id=647501.728168

Dellarocas, C. (2010). Designing reputation systems for the social web. *paperssrncom*, *51*(3), 33—38. Available from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1624697

Dibbell, J. (1993, December). A rape in cyberspace, or how an evil clown, a haitian trickster spirit, two wizards, and a cast of dozens turned a database into a society. *Village Voice.* Available from http://www.juliandibbell.com/articles/a-rape-in-cyberspace/

Dickinson, D. (2004, November). An architecture for spam regulation. *Federal Communications Law Journal*, *57*(1), 129-160. Available from http://law.indiana.edu/fclj/pubs/v57/no1/Dickinson.pdf

DiMicco, J. M., & Millen, D. R. (2007). Identity management: Multiple presentations of self in facebook. In *Proceedings of the 2007 International ACM Conference on Supporting Group Work* (pp. 383–386). New York, NY, USA: ACM. Available from http://doi.acm.org/10.1145/1316624.1316682

Dingledine, R., & Murdoch, S. (2009, March). *Performance Improvements on Tor or, Why Tor is slow and what we're going to do about it* (Tech. Rep.). Walpole, MA: The Tor Project. Available from http://www.torproject.org/press/presskit/2009-03-11-performance.pdf

Doctorow, C. (2007, April). Why online games are dictatorships. *Information Week.*

Available from http://www.informationweek.com/news/199100026

Donath, J. S. (1999). Identity and deception in the virtual community. In M. A. Smith & P. Kollock (Eds.), *Communities in Cyberspace* (p. 29-59). New York, N.Y.: Routledge. Available from http://smg.media.mit.edu/people/judith/Identity/IdentityDeception.html

Donath, J. S., & Boyd, D. (2004, October). Public displays of connection. *BT Technology Journal*, *22*, 71–82. Available from http://dl.acm.org/citation.cfm?id=1031314.1031348

Douglas, K., & McGarty, C. (2001). Identifiability and self-presentation: Computer-mediated communication and intergroup interaction. *British Journal of Social Psychology*, *40*(Pt 3), 399-416.

Dwork, C., & Naor, M. (1993, August). Pricing via Processing or Combatting Junk Mail. In *Proceedings of the 12th Annual International Cryptology Conference on Advances in Cryptology* (p. 139-147). London, UK: Springer-Verlag. Available from http://www.wisdom.weizmann.ac.il/~naor/PAPERS/pvp.ps

Ellison, N., Heino, R., & Gibbs, J. (2006). Managing Impressions Online: Self-Presentation Processes in the Online Dating Environment. *Journal of Computer-Mediated Communication*, *11*(2), 415–441. Available from http://jcmc.indiana.edu/vol11/issue2/ellison.html

Farkas, C., Ziegler, G., Meretei, A., & Lörincz, A. (2002). Anonymity and accountability in self-organizing electronic communities. In *Proceedings of the 2002 ACM workshop on Privacy in the Electronic Society* (pp. 81–90). New York, NY, USA: ACM. Available from http://doi.acm.org/10.1145/644527.644536

Fish, E. (2009). Is Internet Censorship Compatible with Democracy? Legal Restrictions of Online Speech in South Korea. *Asia Pacific Journal on Human Rights and the Law*, *10*(2), 43–96. Available from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1489621

Florencio, D., & Herley, C. (2007). A large-scale study of web password habits. In *Proceedings of the 16th international conference on World Wide Web* (pp. 657–666). New York, NY, USA: ACM. Available from http://doi.acm.org/10.1145/1242572.1242661

Forte, A., & Bruckman, A. (2008). Scaling Consensus: Increasing Decentralization in Wikipedia Governance. In *Proceedings of the Proceedings of the 41st Annual Hawaii International Conference on System Sciences.* Washington, DC, USA: IEEE Computer Society. Available from http://dx.doi.org/10.1109/HICSS.2008.383

Friedman, E., Resnick, P., & Sami, R. (2007). Manipulation-resistant reputation systems. In N. Nisan, T. Roughgarden, E. Tardos, & V. Vazirani (Eds.), *Algorithmic game theory.* Cambridge: Cambridge University Press. Available from http://www.eecs.harvard.edu/cs286r/archived/fall09/papers/repchapter-post.pdf

Ghastin, J. (2009, October 14). Responding to abuse reports more effectively. *Facebook Blog.* Available from http://www.facebook.com/blog.php?post=144628037130

Gibbs, J., Ellison, N., & Heino, R. (2006). Self-Presentation in Online Personals: The

Role of Anticipated Future Interaction, Self-Disclosure, and Perceived Success in Internet Dating. *Communications Research*, *33*(2), 152–177. Available from http://www.msu.edu/~nellison/Gibbs_Ellison_Heino_2006.pdf

Glanz, J., & Markoff, J. (2011, June 12). U.S. Underwrites Internet Detour Around Censors. *The New York Times*. Available from http://www.nytimes.com/2011/06/12/world/12internet.html

Goffman, E. (1959). *The presentation of self in everyday life.* Garden City N.Y.: Doubleday.

Goodman, J., Heckerman, D., & Rounthwaite, R. (2005). Stopping Spam. *Scientific American*, *292*(4), 42–49.

Gross, R., & Acquisti, A. (2005). Information revelation and privacy in online social networks. In *Proceedings of the 2005 acm workshop on privacy in the electronic society* (pp. 71–80). New York, NY, USA: ACM. Available from http://doi.acm.org/10.1145/1102199.1102214

Hansell, S. (2006, February 5). AOL and Yahoo put price on e-mail. *The New York Times*. Available from http://www.nytimes.com/2006/02/05/technology/05iht-email.html

Harris, L., & Llanso, E. J. (2011, June 2). Facebook Age and Anonymity: Civility vs. Freedom of Speech. *ABC News*. Available from http://abcnews.go.com/story?id=13735349\#.T48smjJYuWW

Hayes, B. (1990). Anonymous one-time signatures and flexible untraceable electronic cash. In *Proceedings of the International Conference on Cryptology* (pp. 294–305). New York, NY, USA: Springer-Verlag New York, Inc. Available from http://dl.acm.org/citation.cfm?id=101539.101663

He, F., Leon, P., Luan, Y., & Sun, K. (n.d.). *Future internet based on identification.* Available from http://www.andrew.cmu.edu/user/pgl/FIBI.pdf

Helman, I. (2009). Spam-A-Lot: The States' Crusade Against Unsolicited Email in Light of the CAN-SPAM Act and the Overbreadth Doctrine. *Boston College Law Review*, *50*(5), 1525. Available from http://lawdigitalcommons.bc.edu/bclr/vol50/iss5/10/

Holahan, C. (2006, November). The Dark Side of Second Life. *Business Week*. Available from http://www.businessweek.com/technology/content/nov2006/tc20061121_727243.htm

Johnson, D. R., Crawford, S. P., & Palfrey, J. G. (2004). The Accountable Net: Peer Production of Internet Governance. *Virginia Journal of Law & Technology*, *9*, 2–33. Available from www.vjolt.net/vol9/issue3/v9i3_a09-Palfrey.pdf

Jøsang, A., & Pope, S. (2005, May). User-centric identity management. In A. Clark (Ed.), *Proceedings of AusCERT Asia Pacific Information Technology Security Conference.* Brisbane, Australia: AusCERT. Available from http://folk.uio.no/josang/papers/JP2005-AusCERT.pdf

Judge, P., Alperovitch, D., & Yang, W. (2005, June). Understanding and reversing the profit model of spam. In *Fourth Workshop on the Economics of Information Security.* Cambridge, MA: WEIS '05. Available from http://www.mcafee.com/us/resources/white-papers/wp-understanding-and-reversing-profit-model-spam.pdf

Kanich, C., Kreibich, C., Levchenko, K., Enright, B., Voelker, G. M., Paxson, V., et al. (2008). Spamalytics: An Empirical Analysis of Spam Marketing Conversion. In *Proceedings of the 15th ACM Conference on Computer and Communications Security* (p. 3-14). New York, NY, USA: ACM. Available from http://www.icsi.berkeley.edu/pubs/networking/2008-ccs-spamalytics.pdf

Kelly, K. (2006). *Dangerous Ideas: More Anonymity Is Good.* Available from http://edge.org/q2006/q06_4.html

Kilian, J., & Petrank, E. (1997). Identity escrow. *In Advances in Cryptology — CRYPTO '98*, *1642*, 169–185. Available from http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.21.6420

Kludt, A. (2011, July 22). Yelp Refuses to Remove Fake Review of Anella. *NY Eater Blog.* Available from http://ny.eater.com/archives/2011/07/yelp_refuses_to_remove_fake_review_of_anella.php

Knake, R. (2010, July). *Untangling Attribution: Moving to Accountability in Cyberspace.* Available from http://www.cfr.org/publication/22630/untangling_attribution.html (Testimony before the House of Representatives Committee on Science and Technology)

Koomey, J., Alstyne, M. van, & Brynjolfsson, E. (2007, September 6). You've Got Spam. *Wall Street Journal*, A16. Available from http://www.lbl.gov/today/2007/Sep/07-Fri/kommey-jump.pdf

Kormann, D., & Rubin, A. D. (2000). Risks of the Passport Single Sign-on Protocol. *Computer Networks*, *33*, 51–58. Available from http://avirubin.com/passport.html

Kraut, R., Sunder, S., Telang, R., & Morris, J. (2005). Pricing electronic mail to solve the problem of spam. *Human-Computer Interaction*, *20*, 195-223. Available from http://www.heinz.cmu.edu/~rtelang/Kraut05-PricingEmail.pdf

Lampe, C., Ellison, N., & Steinfield, C. (2006). A face(book) in the crowd: Social searching vs. social browsing. In *Proceedings of the 2006 20th anniversary conference on computer supported cooperative work* (pp. 167–170). New York, NY, USA: ACM. Available from http://doi.acm.org/10.1145/1180875.1180901

Landwehr, C. E. (2009, May). A National Goal for Cyberspace: Create an Open, Accountable Internet. *IEEE Security and Privacy*, *7*, 3–4. Available from http://dl.acm.org/citation.cfm?id=1591889.1592167

Langevin, J. R., McCaul, M. T., Charney, S., Raduege, H., & Lewis, J. (2008). *Securing Cyberspace for the 44th Presidency* (Tech. Rep.). Washington, DC: Center for Strategic and International Studies. Available from csis.org/files/media/csis/pubs/081208_securingcyberspace_44.pdf

Lessig, L. (2006). *Code: And Other Laws of Cyberspace, Version 2.0.* New York, N.Y.: Basic Books. Available from http://codev2.cc/

Levchenko, K., Chachra, N., Enright, B., Felegyhazi, M., Grier, C., Halvorson, T., et al. (2011, May). Click Trajectories: End-to-End Analysis of the Spam Value Chain. In *Proceedings of the 32nd Annual Symposium on Security and Privacy* (pp. 431–446). Oakland, CA: IEEE. Available from http://cseweb.ucsd.edu/~savage/papers/Oakland11.pdf

Levmore, S. X. (2010). The Internet's Anonymity Problem. In S. X. Levmore & M. C. Nussbaum (Eds.), *The Offensive Internet: Speech, Privacy, and Reputation.* Cambridge Mass.: Harvard University Press.

Lieber, R. (2012, March 2). At Angie's List, the Reviews are Real (So Is Angie). *The New York Times.* Available from http://www.nytimes.com/2012/03/03/your-money/at-angies-list-the-reviews-are-real.html

Lowe, L. (2010, March 18). Yelp's review filter explained. *Yelp Blog.* Available from http://officialblog.yelp.com/2010/03/yelp-review-filter-explained.html

Luca, M. (2011, September 16). Reviews, Reputation, and Revenue: The Case of Yelp.com. *Harvard Business School working paper.* Available from http://www.hbs.edu/research/pdf/12-016.pdf

McConnell, M. (2010, February 28). How to win the cyber-war we're losing. *The Washington Post*, B01. Available from http://www.washingtonpost.com/wp-dyn/content/article/2010/02/25/AR2010022502493.html

McCracken, H. (2011, September 11). Google+'s real-name policy: Identity vs. anonymity. *Time.* Available from http://www.time.com/time/business/article/0,8599,2094409,00.html

Micali, S. (1993). *Fair cryptosystems* (Tech. Rep.). Cambridge, MA: Massachusetts Institute of Technology.

Miller, C. C. (2012, January 23). In a Switch, Google Plus Now Allows Pseudonyms. *The New York Times.* Available from http://bits.blogs.nytimes.com/2012/01/23/in-a-switch-google-plus-now-allows-pseudonyms/

Mirkovic, J., & Reiher, P. (2008). Building accountability into the future Internet. In *Proceedings of the 4th Workshop on Secure Network Protocols* (pp. 45–51). Orlando, FL: NPSEC 2008. Available from http://www.isi.edu/~mirkovic/publications/npsec.pdf

Mnookin, J. L. (1996, June). Virtual(ly) Law: The Emergence of Law in LambdaMOO. *Journal of Computer-Mediated Communication*, *2*(1). Available from http://jcmc.indiana.edu/vol2/issue1/lambda.html

Mukhamedov, A., & Ryan, M. D. (2005). On Anonymity with Identity Escrow. In T. Dimitrakos, F. Martinelli, P. Y. A. Ryan, & S. A. Schneider (Eds.), *Formal Aspects in Security and Trust, Third International Works, FAST 2005, Newcastle upon Tyne, UK, July 18-19, 2005, Revised Selected Papers* (Vol. 3866, pp. 235–243). New York, N.Y.: Springer. Available from http://www.cs.bham.ac.uk/~mdr/research/papers/pdf/05-fast.pdf

Narayanan, A., & Shmatikov, V. (2008). Robust De-anonymization of Large Sparse Datasets. In *Proceedings of the 2008 IEEE Symposium on Security and Privacy* (pp. 111–125). Washington, DC, USA: IEEE Computer Society. Available from http://dx.doi.org/10.1109/SP.2008.33

National Strategy for Trusted Identities in Cyberspace. (2011, April). Available from http://www.whitehouse.gov/sites/default/files/rss_viewer/NSTICstrategy_041511.pdf

Newell, P. (2006). Taking Accountability into Account: The Debate So Far. In P. Newell & J. Wheeler (Eds.), *Rights, Resources and the Politics of Account-*

ability (pp. 37–58). London, UK: Zed Books.

Niculescu, M. F., & Wu, D. (2011, May). *When should software firms commercialize new products via freemium business models?* Available from http://rady.ucsd.edu/faculty/seminars/2011/papers/Niculescu.pdf (Working Paper, Georgia Institute of Technology)

Odlyzko, A. M. (2003). The case against micropayments. In R. N. Wright (Ed.), *Financial cryptography '03* (pp. 77–83). Guadeloupe, French West Indies: Springer. Available from http://www.dtc.umn.edu/~odlyzko/doc/case.against.micropayments.pdf

Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011). Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1, pp. 309–319). Stroudsburg, PA, USA: Association for Computational Linguistics. Available from http://dl.acm.org/citation.cfm?id=2002472.2002512

Penenberg, A. (2009). *Viral Loop: From Facebook to Twitter, how today's smartest businesses grow themselves.* New York, N.Y.: Hyperion.

Pfanner, E. (2011, September 4). Naming Names on the Internet. *The New York Times.* Available from http://www.nytimes.com/2011/09/05/technology/naming-names-on-the-internet.html

Pokin, S. (2007, November 16). 'My Space' hoax ends with suicide of Dardenne Prairie teen. *St. Louis Post-Dispatch.* Available from http://www.stltoday.com/suburban-journals/stcharles/news/stevepokin/my-space-hoax-ends-with-suicide-of-dardenne-prairie-teen/article_0304c09a-ab32-5931-9bb3-210a5d5dbd58.html

Priem, B., Leenes, R., Kosta, E., & Kuczerawy, A. (2011). The Identity Landscape. In J. Camenisch, D. Sommer, & R. Leenes (Eds.), *Digital Privacy - PRIME* (pp. 33–51). Berlin, Heidelberg: Springer-Verlag.

Pujol, N. (2010). Freemium: Attributes of an Emerging Business Model. *Social Science Research Network (1718663).* Available from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1718663

Qiu, W., Chen, K., & Gu, D. (2002). A New Offline Privacy Protecting E-cash System with Revokable Anonymity. In *Proceedings of the 5th International Conference on Information Security* (pp. 177–190). London, UK, UK: Springer-Verlag. Available from http://dl.acm.org/citation.cfm?id=648026.744518

Raynes-Goldie, K. (2010, January). Aliases, creeping and wall cleaning: Understanding privacy in the age of Facebook. *First Monday*, *15*(1). Available from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2775/2432

Reid, E. (1999). Hierarchy and power: Social control in cyberspace. In M. A. Smith & P. Kollock (Eds.), *Communities in Cyberspace* (pp. 107–134). London, UK: Routledge.

Reno v. American Civil Liberties Union. (1997). 521 U.S. 844, aff'g American Civil Liberties Union v. Reno, 929 F. Supp. 824 (3d Cir. 1996). Available from http://legal.web.aol.com/decisions/dldecen/reno.html

Resnick, P., Kuwabara, K., Zeckhauser, R., & Friedman, E. (2000, December). Reputation systems. *Commun. ACM*, *43*(12), 45–48. Available from http://doi.acm.org/10.1145/355112.355122

Resnick, P., Zeckhauser, R., Swanson, J., & Lockwood, K. (2003). The Value of Reputation on eBay: A Controlled Experiment. *Experimental Economics*, *9*, 79–101. Available from http://dx.doi.org/10.2139/ssrn.385206

R v. Hampson. QCA 132; BC201104379. (2011, 21 June). Supreme Court of Queensland, Court of Appeal - Judge(s) Muir, White JJA and Daubney, J.

Schedler, A. (1999). Conceptualizing Accountability. In A. Schedler, L. Diamond, & M. Plattner (Eds.), *The Self-Restraining State: Power and Accountability in New Democracies* (pp. 13–28). Boulder, CO: Lynne Rienner Publishers.

Schiano, D. J., & White, S. (1998). The first noble truth of Cyberspace: People are people (even when they MOO). In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 352–359). New York, NY, USA: ACM Press/Addison-Wesley Publishing Co. Available from http://dx.doi.org/10.1145/274644.274693

Schneider, F. B. (2009). Accountability for Perfection. *IEEE Security & Privacy*, *7*(2), 3-4. Available from http://www.cs.cornell.edu/fbs/publications/IEEEspEdtrl.March2009.pdf

Schneier, B. (2006, January 12,). Anonymity Won't Kill the Internet. *Wired*. Available from http://www.wired.com/politics/security/commentary/securitymatters/2006/01/70000?currentPage=all

Schneier, B. (2008, January). Security vs. Privacy. *Schneier on Security Blog*. Available from http://www.schneier.com/blog/archives/2008/01/security_vs_pri.html

Segal, D. (2011, May 21). A rave, a pan, or just a fake? *The New York Times*. Available from http://www.nytimes.com/2011/05/22/your-money/22haggler.html

Smith, A. D. (1999). Problems of conflict management in virtual communities. In M. A. Smith & P. Kollock (Eds.), *Communities in Cyberspace* (pp. 134–166). London, UK: Routledge.

Solms, S. von, & Naccache, D. (1992, October). On blind signatures and perfect crimes. *Comput. Secur.*, *11*(6), 581–583. Available from http://dx.doi.org/10.1016/0167-4048(92)90193-U

Soma, J., Singer, P., & Hurd, J. (2008). Spam Still Pays: The Failure of the CAN-SPAM Act of 2003 and Proposed Legal Solutions. *Harvard Journal on Legislation*, *45*, 165.

Sorkin, D. E. (2001). Technical and legal approaches to unsolicited electronic mail. *U.S.F. L. Rev.*, *35*, 325-384. Available from http://www.sorkin.org/articles/usf.pdf

Springen, K. (2009, November). Vetting Vine Voices. *Publishers Weekly*. Available from http://www.publishersweekly.com/pw/print/20091109/25966-vetting-vine-voices-.html

Sproull, L. (2011). Prosocial Behavior on the Net. *Daedalus*, *140*(4), 140—153.

Stadler, M., Piveteau, J.-M., & Camenisch, J. (1995). Fair blind signatures. In *Pro-*

ceedings of the 14th annual international conference on theory and application of cryptographic techniques (pp. 209–219). Berlin, Heidelberg: Springer-Verlag. Available from http://dl.acm.org/citation.cfm?id=1755009.1755032

Steinhauer, J. (2008, November 21). Woman who posed as boy testifies in case that ended in suicide of 13-year-old. *The New York Times.* Available from http://www.nytimes.com/2008/11/21/us/21myspace.html

Stelter, B. (2008, November 27). Guilty verdict in cyberbullying case provokes questions over online identity. *The New York Times.* Available from http://www.nytimes.com/2008/11/28/us/28internet.html

Streitfeld, D. (2011, August 19). In a Race to Out-Rave, 5-Star Web Reviews Go for $5. *The New York Times.* Available from http://www.nytimes.com/2011/08/20/technology/finding-fake-reviews-online.html

Streitfeld, D. (2012, January 26). For $2 a Star, an Online Retailer gets 5-Star Product Reviews. *The New York Times.* Available from http://www.nytimes.com/2012/01/27/technology/for-2-a-star-a-retailer-gets-5-star-reviews.html

Suler, J. R. (2004). The Online Disinhibition Effect. *Cyberpsychology and Behavior*, *7*, 321–326. Available from http://users.rider.edu/~suler/psycyber/disinhibit.html

Sun, S.-T., Boshmaf, Y., Hawkey, K., & Beznosov, K. (2010). A billion keys, but few locks: the crisis of web single sign-on. In *Proceedings of the 2010 workshop on new security paradigms* (pp. 61–72). New York, NY, USA: ACM. Available from http://doi.acm.org/10.1145/1900546.1900556

Sun, S.-T., Pospisil, E., Muslukhov, I., Dindar, N., Hawkey, K., & Beznosov, K. (2011). What makes users refuse web single sign-on?: An empirical investigation of OpenID. In *Proceedings of the Seventh Symposium on Usable Privacy and Security* (pp. 4:1–4:20). New York, NY, USA: ACM. Available from http://doi.acm.org/10.1145/2078827.2078833

Tabuchi, H. (2011, January 9). Facebook Wins Relatively Few Friends in Japan. *The New York Times.* Available from http://www.nytimes.com/2011/01/10/technology/10facebook.html

Taylor, T. L. (2006, September). Beyond Management: Considering Participatory Design and Governance in Player Culture. *First Monday.* Available from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1611/1526

Tresca, M. J. (1998). *The Impact of Anonymity on Disinhibitive Behavior Through Computer-Mediated Communication.* Unpublished master's thesis, Michigan State University. Available from http://www.msu.edu/user/trescami/thesis.htm

U.S. v. Lori Drew. (2009). 259 F.R.D. 449. C.D. Cal. Available from http://volokh.com/files/LoriDrew.pdf

Veneziani, V. (2006, September 8). Metaverse breached: Second Life customer database hacked. *TechCrunch.* Available from http://techcrunch.com/2006/09/08/metaverse-breached-second-life-customer-database-hacked/

Wang, R., Chen, S., & Wan, X. (2012, May). Signing Me onto Your Accounts through

Facebook and Google: a Traffic-Guided Security Study of Commercially Deployed Single-Sign-On Web Services. In *Proceedings of the IEEE Symposium on Security and Privacy.* Oakland, CA: IEEE Computer Society. Available from `http://research.microsoft.com/pubs/160659/websso-final.pdf`

Weise, K. (2011, September 29). A lie detector test for online reviewers. *Bloomberg Businessweek.* Available from `http://www.businessweek.com/magazine/a-lie-detector-test-for-online-reviewers-09292011.html`

Wheaton, S. (2007, February 28). (Mis)adventures in Cyberspace. *The New York Times.* Available from `http://thecaucus.blogs.nytimes.com/2007/02/28/misadventures-in-cyberspace/`

Wright, L. (2008, January 21). The Spymaster. *The New Yorker.* Available from `http://www.newyorker.com/reporting/2008/01/21/080121fa_fact_wright`

Zahorsky, I. (2011, August). Tor, Anonymity, and the Arab Spring: An Interview with Jacob Appelbaum. *Peace & Conflict Monitor.* Available from `http://www.monitor.upeace.org/innerpg.cfm?id_article=816`

Zhao, S., Grasmuck, S., & Martin, J. (2008). Identity construction on Facebook: Digital empowerment in anchored relationships. *Computers in Human Behavior, 24*(5), 1816 - 1836. Available from `http://astro.temple.edu/~bzhao001/Identity%20Construction%20on%20Facebook.pdf`

Zhuo, J. (2010, November 29). Where Anonymity Breeds Contempt. *The New York Times.* Available from `http://www.nytimes.com/2010/11/30/opinion/30zhuo.html`

Zittrain, J. (2009). *The Future of the Internet—And How to Stop It.* New Haven, Conn.: Yale University Press. Available from `http://futureoftheinternet.org/download`