An Uneasy Relationship: Cyber Security Information Sharing, Communications Privacy, and the Boundaries of the Firm

Aaron J. Burstein*
Samuelson Law, Technology & Public Policy Clinic
Berkeley Center for Law & Technology
School of Law (Boalt Hall)
University of California, Berkeley
aburstein@law.berkeley.edu

March 1, 2007

Abstract

Cross-organizational sharing of network traffics data has the potential to provide researchers and network operators with much needed insight to develop defenses against highly distributed attacks on network infrastructure. Previous work has focused on the technical and economic aspects of allowing organizations to share network traffic data, or allowing a centralized investigator to do so; but the equally fundamental concerns arising from the legal protection of communications data has received little detailed attention in this context. This paper seeks to fill that gap.

As a prologue to further economic analysis of this problem, this paper analyzes how communications privacy law would affect the deployment of two recent proposals for cross-organizational sharing of network traffic data. The central issue in both proposals is that many communications records are protected from disclosure in the United States by the Stored Communications Act (SCA). Though both proposals include steps to protect the confidentiality of shared data, the SCA would likely bar commercial Internet service providers (ISPs) from sharing data under either system. Thus, this paper lays the groundwork for further study of the kinds of institutions that are necessary to support cross-organizational communications data sharing. Clarifying the legal framework will allow a more realistic assessment of the returns that contributors of network traffic data may expect.

*TRUST Research Fellow. This work was supported in part by TRUST (The Team for Research in Ubiquitous Secure Technology), which receives support from the National Science Foundation (NSF award number CCF-0424422). I am grateful to Deirdre Mulligan sharing many key ideas on this topic and to Joseph Lorenzo Hall and Jen King for reviewing a draft of this paper.

1 Introduction

Internet-scale attacks have emerged as a powerful tool for committing financial crimes, and they present a serious threat to the global information infrastructure itself [30]. Criminals use "botnets"—networks of compromised computers under central control—to collect and exchange personal and financial information, as well as to move money around in a manner that is very difficult to trace [20]. Botnets have also become the tool of choice for launching distributed denial of service (DDOS) attacks, since botnets offer widely distributed bandwith and are difficult to trace to a single source [19, 18]. Similarly, traffic from worms and viruses can be difficult to detect from a single network monitor [31]. Effective detection and forensic analysis of these attacks, as well as the development of better defenses, depend on different organizations being able to share network traffic data [28].

Despite the recognized need to share network data in order to improve cyber security, actual cooperation among organizations that possess this data has been slow to emerge [23]. As [7] has pointed out, cyber security problems have the characteristics of the tragedy of the commons: ISPs do not control the hardware and software in their users' homes and businesses. Software vendors do not pay for the damage that exploits of their products cause to data, hardware, or networks, or for the role that they play in facilitating increasingly sophisticated network attacks. Users do not pay for the damage that their vulnerable computers cause. Moreover, ownership of the pieces of the Internet, from end-users to local ISPs to backbone operators, is highly fragmented, making data collection for cyber security research a challenging technical and economic problem.

Add to this list of difficulties legal protection for the privacy of network traffic data. Privacy protections for these data depend on who collects the data, the legal (rather than technical) classification of what the data represent, and the identity of any recipient of disclosed data. In recent years a number of cyber security researchers have proposed ways of sharing that are technologically feasible, would serve researchers' needs, and at least recognize the need to protect users' privacy rights. But an important question remains to be explored: What exactly are the legal obstacles to implementing proposals? Furthermore, assuming that there are no changes in the easing restrictions on disclosing communications data, what kinds of organizational structures would permit some, and how would these structures affect the incentives of firms to share information?

This paper answers these questions for two specific network information-sharing proposals, but the results are applicable to the broader debate about sharing security information. Particularly, I argue that anonymizing data, though it may protect privacy, does not fully address statutorily created rights of privacy in communications data. I find that these two information-sharing models would face significant legal hurdles unless implementations of those systems include the necessary limitations on information disclosure. I discuss a few approaches to these disclosure limitations and point out how those approaches might alter the economic landscape of the problem.

Part 2 summarizes work from the computer science and economics literature regarding the cyber security information sharing problem. Part 3 summarizes the primary impediment under United States law—the Stored Communications Act (SCA)—and briefly discusses other legal considerations. In Part 4 I analyze the issues that pertain specifically to proposals to share information for defending against Internet-scale attacks. Part 5 presents some suggestions for addressing these legal issues. I conclude and present ideas for future work in Part 6.

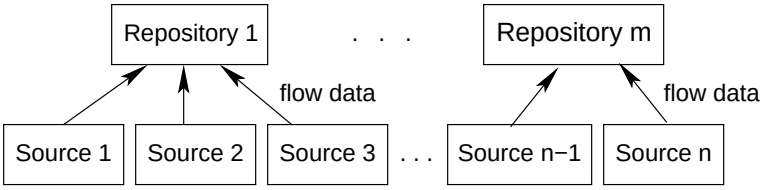

Figure 1: Communications data flow from sources to repositories in the Horizontal Model.

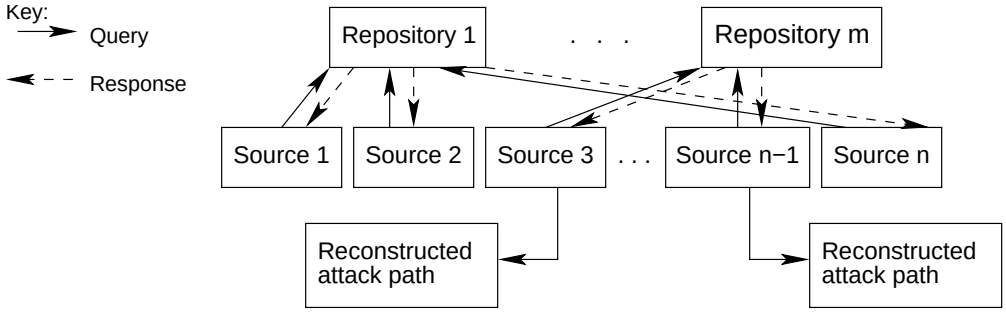

Figure 2: A possible query-and-response pattern in the Horizontal Model.

2 Related Work

2.1 Computer Science

Two recent approaches to sharing network traffic data in the network defense context span a considerable scientific and legal range. For both models, it is important for my purposes to state clearly who collects data, what data they collect, what data they share, and with whom they share it.

The first system that I summarize was described by [26]; I call it the Horizontal Model, because it would allow data contributors to query repositories containing data contributed by other entities.[1] The authors suggest that traffic data should be collected at the administrative domain [15] or autonomous system level [16]. Specifically, the authors estimate that if 50 of the most highly connected autonomous systems participated in the Horizontal Model, the resulting system would cover approximately 90% of Internet paths. Basically, the Horizontal Model would have Internet domains with a high degree of connectivity collect "flow" records. These records would be relatively terse, noting only the direction of communication, source and destination Internet Protocol (IP) addresses, and timestamps marking the beginning and end of the communication flow.

Under the Horizontal Model, data collectors would share flow data with other collectors. The collectors would not necessarily share raw flow records, but rather would store them in repositories and allow other collectors to query the database. The details of how participants in the Audit System would reconstruct attacks from this system of distributed traffic flow databases need not concern us here. Rather, the important element of the Horizontal Model is what it would allow one domain to disclose to another: flow information. In other words, source and destination IP ad-

[1] [26] suggests that an implementation of the Horizontal Model should produce evidence with accuracy and chain-of-custody protocols that are sufficient to make the data useful to law enforcement officials. I focus on whether it is feasible to make such data available to researchers and leave the discussion of how cross-organizational information sharing should relate to law enforcement for another day.

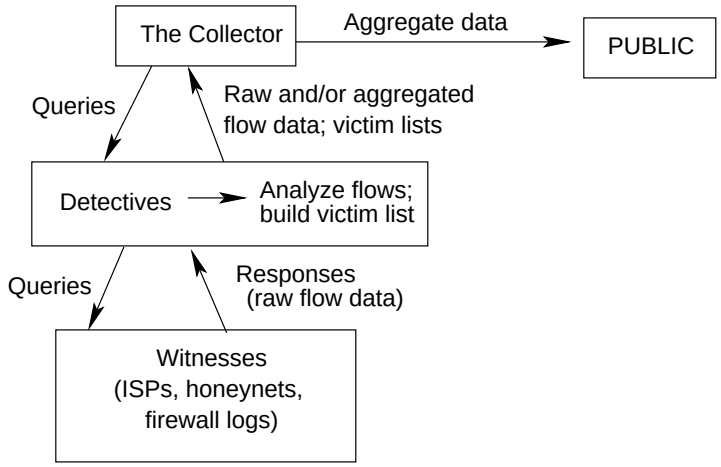

Figure 3: A schematic diagram of communication records flows under the Vertical Model.

dresses relating to users within a domain would be transferred. This exchange creates far-reaching consequences under the SCA.

The second model, which I call the Vertical Model, is described in [6]. This model defines three entities: witnesses, detectives, and the collector. The first type, witnesses, are the most numerous and are members of an open set. They are not trusted and serve only to collect traffic data. For example, an ISP that logs connection data from its customers could serve as a witness. The second kind of entity is a detective, which queries witnesses for the presence of certain patterns in witnesses' traffic logs. As described in [6], the detectives are a closed set of trusted entities. The detectives send their analyses to the third type of entity, the collector, a "network Interpol" that aggregates (and might further analyze) and publishes detectives' reports.

2.2 Economics

Economic studies of security information-sharing have largely focused on the problem of determining when it would be profitable for firms to share information about vulnerabilities in their own computer systems or networks. [14] points out that the benefits to firms of sharing information depend heavily upon what kinds of information are exchanged, what kind of competition is present in an industry, and whether participating firms make the shared information public. A more recent analysis finds that firms are more likely to share security information in large industries populated by large firms [13] . This sharing can lead to two positive effects for firms that contribute information: a direct benefit resulting from increased demand for their services; and a strategic effect resulting from an ability to set higher prices. Nonetheless, it is worth noting that a number of projects collect network security information primarily from volunteer sources who stand to gain little direct benefit from their contributions [11, 21].

For institutions focused on research, such as universities and national labs, the effects described above might be less important. Instead, these institutions (and the researchers whom they employ) would likely want to share information in order to gain access to other firms' information, if such a trade-off is required. Research institutions might also be good candidates for providing critical mass and infrastructure for cyber security information-sharing. As explained in detail in Parts 4

and 5, however, the presence of government-affiliated researchers vastly complicates the sharing of network communications data.

Governments also sometimes promote security information sharing, as the U.S. Departments of Defense and Homeland Security do by funding the CERT Coordination Center [10], [8]. The Department of Homeland Security is also providing funding for a Protected Repository for the Defense of Infrastructure Against Cyber Threats (PREDICT), but this project has yet to become operational [24]. For the reasons discussed below, legal restrictions on the government's access to communications data play a large role in preventing the government from participating in ventures that share this kind of information.

Aside from acknowledging that some kinds of information sharing may raise antitrust issues, economic studies of information sharing have assumed that the participants have more or less free rein in deciding how and what to share [29, 27, 12]. A contribution of this paper is to examine this assumption in the context of sharing communications data. As I explain below, U.S. law imposes considerable constraints on the extent to which firms may share these data.[2] Other costs that might discourage firms from sharing network data include exposing security vulnerabilities in their own systems and the risk of lawsuits from users who allege that a firm has violated its obligation to keep such data confidential.

3 The Problem of Sharing Communications Data

The basic problem that I address is to identify and analyze the salient legal constraints on sharing the kinds of communications data that cyber security researchers, as typified by the Horizontal Model and the Vertical Model, would find useful. The primary source of protection for communications data in the United States is a statute known as the Electronic Communications Privacy Act (ECPA). The ECPA has three principal parts: the Wiretap Act [5], the Pen Register statute [3], and the Stored Communications Act (SCA) [4].

The SCA provides the major obstacle to sharing communications data in the manner described by the two Models. In addition to summarizing the SCA, I briefly discuss the other parts of the ECPA to justify eliminating them from further discussion here.

- **SCA:** The pertinent part of the SCA (18 U.S.C. § 2702(a)(3)) prohibits an "electronic communications service provider to the public" from "knowingly divulg[ing] a record or other information pertaining to a subscriber to or customer of such service . . . to any governmental entity." The records to which this section refers are commonly labeled "non-content" or "addressing" records. These records include a broad array of information: a subscriber's name and address as well as logs that record the IP addresses to which a subscriber connects. This portion of the SCA has some important limitations that are worth noting.
  - **Electronic communication service** is defined to mean "any service which provides to users thereof the ability to send or receive wire or electronic communications." ISPs are included within this definition, as are e-mail providers and instant messaging services. Because both the Horizontal Model and the Vertical Model focus on data that ISPs are

[2]Communications data protection laws vary considerably from country to country. Extending this analysis to include these international considerations is one potentially fertile area for future work.

likely to have (and other types of electronic communications services are not), I shall refer to ISPs rather than electronic communication services for the remainder of this paper.

– The phrase **"to the public"** significantly limits the scope of the SCA's application; an ISP that does not offer services to the public is not regulated by the SCA. The touchstone for deciding whether a given ISP meets this requirement is whether the ISP requires some kind of special relationship with a user before it will provide service to that user. A commercial ISP offers its service to all users in a geographic area and is clearly an ISP "to the public" and is subject to the SCA.

  By contrast, a private employer that provides Internet connectivity only to its employees is not an ISP "to the public"; the employer-employee relationship is a prerequisite for accessing the Internet via the employer's network.

  There may also be intermediate cases. Public universities, for example, typically offer Internet connectivity to their students, staff, and faculty. A university, however, might also offer temporary access to visitors. Depending on the level of university-related justification necessary to obtain this access, a court might find that the university provides service "to the public." In any event, many institutions that are not subject to the SCA have policies that prohibit them from sharing communications records outside of a select group of officials within the institution.

– Similarly, the qualification that the SCA applies to a **"subscriber or customer"** of the ISP might limit its application. This category is narrower than "user," which is defined elsewhere in the ECPA. This language reflects the conditions at the time that the SCA was enacted, but it may function to limit the scope of the SCA's application to current technology. For example, it is unclear whether a user of a community and municipal wireless services is a "subscriber or customer" of those services.

– **"Knowingly divulge":** Although the SCA does not define specifically what it means to act "knowingly," a court that examined this issue held that it means that a party was "aware of the nature of the conduct, aware of or possessing a firm belief in the existence of the requisite circumstances and an awareness [sic] of or a firm belief about the substantial certainty of the result." [1].

  The other half of this phrase has received little attention from courts and legal commentators. In most cases involving the SCA, an ISP has shared communications records with a third party that is unrelated to the ISP, aside from the circumstances surrounding the disclosure of information. That is, divulgement is clear when communications records move between independent firms. Conversely, the SCA does not regulate how communications records may be divulged within a firm. Whether any divulgement of records takes place in hybrid forms of organization may present a closer question. This question becomes especially relevant in the analysis in Part 5.

– **Exceptions:** There are two especially relevant exceptions to sharing information under the Models discussed above. First, the SCA only prohibits disclosure of non-content information to a "governmental entity."[3] The SCA does not define "governmental en-

[3]The SCA, however, prohibits disclosure of the stored contents of communications by a covered entity to any third party.

tity," but the term includes more than law enforcement officers and agencies; it may include the employees or agents of any political subdivision of any state or of the federal government. Under such a broad interpretation, a researcher at a state university or a national lab would qualify as a "governmental entity."

The second exception to the SCA that is worth considering here is consent. An ISP that is subject to the SCA may disclose pertaining to a subscriber if that subscriber gives his or her consent. Although ISP terms of use and privacy policy require users to consent to some disclosure of information regarding their Internet use, these policies typically do not go as far as either of the Models would require.

Government agencies, including law enforcement agencies, may compel SCA-regulated entities to disclose information protected under the SCA. Since I am concerned with the problem of voluntary disclosure, I do not discuss the SCA's provisions for compelled disclosure. See [17] for a helpful exposition of these provisions.

- **Liability:** An ISP regulated by the SCA is subject to a civil lawsuit by a "subscriber . . . or other person aggrieved by any violation" of its provisions [25], and the SCA provides that any person who brings suit is entitled to at least $1000 in damages. Given the number of customers who might be "aggrieved" by the disclosure of an ISP's traffic logs, this minimum would act as a stiff deterrent to any SCA-regulated entity's participation in an information-sharing system. In addition, federal and state governmental agencies may be liable for damages, and government employees are may be disciplined for certain violations of the SCA [25, 2].

- **Wiretap Act.** The Wiretap Act prohibits any person from intercepting the contents of the electronic communications of any other person. The Act also prohibits disclosing, as well as using, the contents of an improperly intercepted communication.[4] Thus, the Wiretap Act differs from the SCA in two important ways: it applies to real-time interception rather than access to stored communications or records; and it applies to the contents of communications, rather than the non-content "addressing" information that is the subject of the SCA. The distinction between content and non-content information is not always clear and does not have a convenient technical rule. For example, the To: and From: fields in an e-mail header are considered to be non-content information, even though they are typically carried in the payload section of IP packets [22]. In the case of the Horizontal Model and the Vertical Model, however, the information that would be exchanged quite clearly falls into the non-content category. Thus, I will not consider the Wiretap Act further in this paper.

- **The Pen Register statute.** The Pen Register statute is the non-content counterpart to the Wiretap Act; it regulates real-time collection of communications addressing information.[5] The statute generally prohibits any person from installing or using a device that collects addressing information in real time, though law enforcement officers may do so if they obtain

[4]The Wiretap Act imposes the same prohibitions on the interception, disclosure and use of oral communications and wire communications (i.e., those carried by a telephone line).

[5]The term "pen register" refers to a device that records incoming phone numbers on a circuit-switched telephone network. Outgoing numbers are recorded using a "trap and trace" device. Because the use of both types of devices is regulated by this statute, it is sometime called the "Pen/Trap statute."

a court order. The Pen Register statute also provides an exception for a communication service provider to use pen/trap devices in relation to the "operation, maintenance, and testing" of their services, to protect its "rights or property," or to protect users of the service. This exception permits ISPs to collect the kinds of information required for input into the Models.

To summarize, there are a few distinctions worth highlighting between the SCA, on the one hand, and the Wiretap Act and Pen Register statute on the other. First, the SCA applies only to entities that provide communication services to the public, while the Wiretap Act and Pen Register statute apply to everyone, including individuals, companies, and the government. Many entities that effectively serve as ISPs for a wide swath of users do not fit this description and thus are not regulated by the SCA. Second, the SCA provides a broad exception that permits disclosure of communication records and contents to any non-governmental entity. Finally, the three statutes that constitute the ECPA are not the only laws that might limit the application of the information-sharing models in practice. Mitigation strategies might implicate state or federal computer abuse statutes. Certain methods for detecting and analyzing communications and malicious code might also implicate traditional copyright law and the Digital Millennium Copyright Act. I address these issues in a separate paper [9].

4 Applying the SCA to the Information Sharing Models

The section analyzes how the Horizontal Model and the Vertical Model interact with the SCA. My approach is to examine whether and how the SCA applies to each entity defined by the two Models, relate this status to the information that each type of entity divulges (or receives), and conclude with a statement of the legal risks that participants in both Models would face.

4.1 Horizontal Model

The first consideration under the Horizontal Model is how to classify in terms of the SCA the entities that provide data. According to [26], the best candidates for data sources would be entities at the administrative domain or autonomous system level. Thus, the sources are likely to be a mixture of entities that provide Internet connectivity. It is likely that some of these sources would be enterprise networks or universities. As discussed in Part 3, these probably are not subject to the SCA. But commercial ISPs, to which the SCA does apply, would probably also be among the data sources. This fact alone would not prevent commercial ISPs from sharing information under the Auditing Model, but it does introduce some complications to this scheme.

4.1.1 Populating the repositories

Specifically, any ISP that is subject to the SCA must avoid divulging network traffic data to a governmental entity. The first risk for divulgement that would violate the SCA arises when a domain sends data to a repository. The precise details of the repository structure have not been specified, but three possibilities are worth considering. One possibility is that the ISP controls the repository, in which case no network traffic data crosses organizational boundaries, and no divulgement occurs. Alternatively, the ISP might send data to a repository controlled by some other organization. In this case the ISP would divulge data to the entity that controls the repository.

If the ISP is subject to the SCA and the repository controller is a governmental entity, then the ISP is probably violating the SCA. For example, if the University of California offered to host datasets from all ISPs in the state—commercial and otherwise—then the commercial ISPs would violate the SCA by sending flow data to the university's repository. On the other hand, a private ISP, such as a large corporation, could provide the repository with data without risking a violation of the SCA. A third possible structure for repositories under the Auditing Model would be to place them under control of some kind of joint venture. This structure might be attractive to participants because it would allow them to share the costs of operating the repository. Again, sending data to the repository involves divulgement; the data move from the ISP to the joint venture. A further question is whether the ISP has divulged data to any other entity—the venture participants, for example—at this point. It is impossible to reach a definitive conclusion without knowing the specific rules governing data transfer from the repository to the joint venture's participants, though I emphasize that this question is only significant to the extent that there is a governmental entity participating in the joint venture.

4.1.2 Responding to queries

Data sources in the Horizontal Model also face the risk that data they provide to a repository will be divulged to a governmental entity in response to queries from parties with access to the repositories. Two conditions must hold in order to find a violation at this stage. First, the data source must be subject to the SCA. Second, the investigator—i.e., the party submitting the query and receiving a response—must be a governmental entity. Recall that this term includes not only law enforcement agencies but also researchers from national labs or state universities.

A potential argument against finding a violation of the SCA under these conditions is that the data source might not know the identity of each potential investigator; or, even if it does, the source does know whether any data that it contributes will be divulged to a particular investigator. This argument is unlikely to relieve the ISP of liability under the SCA. As discussed in Part 3, the state-of-mind requirement in the SCA is that an ISP *knowingly* divulge protected information to a governmental entity. Meeting this standard does not require absolute certainty that the data will end up in a governmental entity's possession, but rather knowledge of the circumstances surrounding divulgement and "substantial certainty" about the result. If an ISP knows, at the time that it contributes data to a repository, that the repository might send data that is responsive to a governmental entity's queries, then it seems highly likely that that source's data will flow to a governmental entity at some point. Given the sensitive nature of the data that the Horizontal Model envisions making available, ISPs would probably examine with great care all plans for sharing data, including the identities of entities that will have access to the repository. In addition, the costs of collecting and transmitting data under this model are likely to be considerable, which also suggests that ISPs will take a close look at an information-sharing scheme based on the Horizontal Model. Both of these considerations make it likely that an ISP would know the identity of the investigator and also that at least some of the data that it provides would be divulged to the investigator.

In summary, though it is likely that only a small fraction of the records that a source contributes will be divulged to the investigator, the likelihood that *some* records will is probably sufficient to meet the SCA's knowledge standard.

4.2 Vertical Model

Like the Horizontal Model, the Vertical Model requires an analysis of the three types of entities it defines: witnesses, detectives, and the collector. Despite the considerable differences in the two Models' proposed information-sharing architectures, the legal issues that they raise are quite similar.

4.2.1 Risks from disclosing to detectives

Like the data sources under the Horizontal Model, the witnesses in the Private Query Model must assess whether they are regulated by the SCA. The Vertical Model places less emphasis than the Horizontal Model on obtaining raw network traffic data from top-level entities; firewall logs and honeynet data—perhaps from address space controlled by a university—could provide data. These kinds of detectives might fall outside the scope of the SCA. Still, the Vertical Model would also take data from ISPs. Each participant in the Vertical Model would have to assess whether it is regulated by the SCA; the relatively decentralized architecture of this model would probably make it difficult for the higher-level entities that receive data from the witnesses to determine whether a given witness is subject to the SCA.

If a witness is subject to the SCA, it must determine whether a detective making a query is a "governmental entity." The Vertical Model envisions that there might be hundreds of detectives, and a witness could receive a query from any one of them. Thus, it would be impossible for a witness to determine in general whether it will disclose data to a governmental entity. I propose some ways of handling this problem in Part 5.

4.2.2 Risks from downstream disclosure by detectives

Witnesses must also consider whether the sharing of data between a detective and the collector might violate the SCA. Again, the threshold question is whether the collector is a governmental entity; if it is, then disclosing network traffic data to it may violate the SCA. Assuming that the collector is some kind of governmental entity and that the witness is also subject to the SCA, the next question concerns the what kinds of data that are provided to the collector.[6] It appears that at least some of the information sent to the collector would be user-level connection data, which is within the purview of the SCA. ([6] states that the collector will "aggregate" data provided via detectives.) Although the detectives would discard much of the uninteresting data, this would not excuse the divulgement of other data.

4.2.3 Risks from disclosures by the collector

Finally, [6] proposes to have the collector make network traffic data available, provided that it obscures the source of each piece of data. This step raises a difficult question under the SCA. I set aside the question of whether a witness would satisfy the SCA's requirement of acting "knowingly" in order to focus on the question of whether obfuscating the source of data, or anonymizing it, would take disclosure out of the SCA's scope.

[6] It is the witnesses, rather than the detectives, that collect the raw data and may be subject to the SCA.

De-identifying data in these ways probably does not make the SCA inapplicable. The most important reason is that the SCA protects communication records from divulgement even if the records are not identified with a customer's name. All that the SCA requires is that a communications record "pertai[n] to" a customer. Moreover, the SCA distinguishes between basic subscriber information—name, address, etc.—and other non-content records held by an ISP. Indeed, the SCA affords greater protection to connection records than to basic subscriber information. This distinction stems from an intent to protect communication records, even if there is no immediate link between the record and a specific, identified person, because the fine-grained detail that they may contain can reveal a great deal of information about a person. This is not say that de-identification would not help to protect individual users' privacy, but it is probably insufficient to remove the data from the broad protection offered by the SCA.

Of course, this line of argument only applies to data whose source is an entity covered by the SCA; other entities do not face this legal obstacle in providing publicly released data. Still, in those cases, de-identifying data is important to protecting individual users' privacy and to providing some security for the source of the data.

5 Suggestions

This section offers some suggestions for refining both Models in order to allow their participants, or to delineate which participants are likely to participate legally. This section also explores how these changes might affect the incentives of firms to participate in a cross-organizational information sharing scheme and presents a few ideas about the kinds of institutions that would allow this sharing to take place under the SCA.

5.1 Horizontal Model

5.1.1 Use data sources that are not subject to the SCA

A crude but effective way of avoiding trouble under the SCA is to use data only from those entities that are not subject to it: administrative domains corresponding to entities that do not offer service to the public. In practice, entities that are subject to the SCA would be unlikely to participate in communication data-sharing under the Horizontal Model for the reasons described in Part 4.1. (But see below for suggestions that might convince these firms to participate.)

Keeping SCA-regulated data sources out of the information-sharing venture might have grave effects on its ability to fulfill the goal of reconstructing Internet-scale attacks. The Horizontal Model suggests using domains with high degrees of connectivity; if a significant number of these domains are subject to the SCA, the data collected within a scheme that excludes these entities might not offer a sufficiently complete picture of Internet host connectivity.

5.1.2 Segregate data from SCA-regulated sources

A more flexible alternative would be to devise an architecture that guaranteed that SCA-regulated entities would send datasets only to repositories controlled by non-governmental entities. This would remove concerns based on the SCA, at least at the point of initial disclosure of the data. Given the relatively small number of data sources and repositories, as well as the extensive work

that would need to go into implementing this model of information sharing, creating this kind of guarantee should be tractable. Further steps that would prevent disclosure that violates the SCA include using internal controls to ensure that any repositories controlled by governmental entities do not receive data from SCA-regulated entities. A policy that prevents one source's data from flowing from a repository to another (potentially governmental) source would also protect data sources from SCA violations.

5.1.3 Restrict queries by governmental entities

A final safeguard would be to limit queries by governmental entities. One way to do this is simply to prevent any governmental entity from querying any repository. In effect, this would require governmental entities to serve exclusively as data sources; they could contribute data but would be unable to extract it. Alternatively, if data from SCA-regulated entities were segregated as discussed above, governmental entities could limit their queries to repositories that hold data from other sources. A slight variation on this idea would include a means for identifying the source of a query as (non)-governmental, which would allow the repository to decided whether to respond to a given query. In any event, considerations arising from the SCA would limit governmental entities' access to data. This, in turn, might reduce their incentives to share their own information. To make any of these approaches administratively feasible would probably require some form of central coordinating body that is responsible for admitting data sources and determining the extent of their permission to query the repositories. It is unclear whether [26] envisions opening repositories to general access, but the considerations discussed here would counsel limiting access to entities that contribute data, or at least to entities whose identities are known and thoroughly understood.

5.2 Vertical Model

The Vertical Model presents a different set of challenges because the set of data contributors—the witnesses—is open. Aside from building the information-sharing system entirely from entities that are not subject to the SCA, there are couple of approaches that would avoid most trouble from this statute.

5.2.1 Label non-governmental detectives

The initial legal risk is that a witness will divulge data to a governmental entity. One way to avoid this problem is for non-governmental detectives to assert that they are, in fact, not a governmental entity. The technical details of this mechanism are beyond the scope of this paper, but perhaps some combination of certificates (to establish identity) and a list of governmental detectives maintained by a trusted host would allow witnesses to make this determination. A witness that is not subject to the SCA could simply ignore this assertion, while other witnesses could refuse to answer queries from a governmental detective.

5.2.2 Label and segregate data for non-governmental detectives

A related suggestion is to allow a witness to specify whether the data that it submits may be shared with a governmental entity. This specification could be part of each record that a witnesses

returns to a detective. This would be a small amount of data, relative to the rest of the record that witnesses return, and so would add little overhead to the system. It might be difficult, however, to convince witnesses that the infrastructure at the detective level would respect the desired limitations disclosure of the data.

5.2.3 Use a non-governmental collector

Making the collector, which sits alone at the top of the Vertical Model's hierarchy, a non-governmental entity would serve the same purpose as discussed in connection with the Horizontal Model's investigator. (See Part 5.1.3.)

Under current law, however, the goal of allowing the collector to publicly release data about attacks might be beyond reach. As discussed above, releasing de-identified data would still put the data source (i.e., the witness) at risk of a violation of the SCA. Restricting disclosure to non-governmental recipients would mitigate this risk, but it would also deprive researchers at government institutions of the benefit of this information.

5.3 Effect of Suggestions on Participants' Incentives

In Part 2.2 I stated that the likely participants in either of these information-sharing schemes would be motivated by a desire to advance research within their own institutions or to improve the state of security on their own networks (or both). In either case, the return for participating in information-sharing would be gaining access to reconstructions of cross-organizational attacks, which an organization's own data contribution would make possible. Unfortunately, the suggestions presented here would require that the information-sharing ventures keep all data contributed by SCA-regulated entities away from governmental entities, including researchers employed by those entities. In effect, governmental entities would contribute data that others could use, but they would only be allowed access to data contributed by entities that are not subject to the SCA. Being denied access to this fairly diverse range of data might lead these institutions to conclude that it is not worthwhile to share communications data with such a broad array of organizations.

6 Conclusions and Future Work

Cross-organizational sharing of non-content communications data holds great promise for advancing cyber security research, particularly detecting and mitigating Internet-scale attacks. At the same time, sharing such information on the scale proposed in the scientific literature presents formidable challenges under U.S. communications privacy law. These legal difficulties are not mere formalities; the models discussed in this paper would involve sharing massive amounts of communications data that are potentially traceable to individual users. The proposals for making these models legally viable would most likely limit the utility of the overall approach, but they would at least allow some work on them to proceed.

A long-term approach that might serve cyber security research without creating undue risk to individual users' privacy would be to create a research exception to the SCA. Creating such a policy would require a closer examination of the privacy risks involved at each stage of information

sharing, as well as a more thorough analysis of the legal and technological means of guarding the confidentiality of shared information.

Another approach would be to explore the creation of joint ventures to allow diverse organizations to contribute data and researchers to the venture. This would be a subtle business, requiring close analysis of the venture's governance, the extent of integration in the venture, and the relationship between the venture and its participants.

Finally, this paper has addressed U.S. law only. Assembling a picture representative of the whole Internet requires data from other countries. An important subject for future work is to determine whether different countries' communications privacy laws conflict and, if so, how they might be reconciled in order to permit the study of Internet-scale attacks.

References

[1] Freedman v. AOL, Inc., 329 F. Supp. 2d 745 (E.D. Va. 2004).

[2] Freedman v. AOL, Inc., 303 F. Supp. 2d 745 (D. Conn. 2004).

[3] Pen register statute. 18 U.S.C. §§ 3121-3127.

[4] Stored communications act. 18 U.S.C. §§ 2701-2712.

[5] Wiretap act. 18 U.S.C. §§ 2510-2522.

[6] Mark Allman, Ethan Blanton, Vern Paxson, and Scott Shenker. Fighting coordinated attackers with cross-organizational information sharing. November 2006.

[7] Ross Anderson. Why information security is hard—an economic perspective. Annual Computer Security Applications Conference, New Orleans, Louisiana, December 2001.

[8] Ross Anderson and Tyler Moore. The economics of information security: A survey and open questions. Fourth bi-annual Conference on the Economics of the Software and Internet Industries, Toulouse, France, January 2007.

[9] Aaron J. Burstein. Toward a culture of cyber security research. *In preparation*, 2007.

[10] CERT. `http://www.cert.org/`, 2007.

[11] DShield. `http://www.dshield.org/`, 2007.

[12] Esther Gal-Or. Information sharing in oligopoly. *Econometrica*, 53(2):329–343, 1985.

[13] Esther Gal-Or and Anindya Ghose. The economic incentives for sharing security information. *Information Systems Research*, 16(2):186–208, 2005.

[14] Larry Gordon, Martin Loeb, and William Lucyshyn. An economics perspective on the sharing of information related to security breaches. First Workshop on the Economics of Information Security, May 2002.

[15] S. Hares and D. Katz. Administrative domains and routing domains: A model for routing in the internet. RFC 1136, December 1989.

[16] J. Hawkinson and T. Bates. Guidelines for creation, selection, and registration of an autonomous system (as). RFC 1996, March 1996.

[17] Orin S. Kerr. A user's guide to the Stored Communications Act—and a legislator's guide to amending it. *George Washington University Law Review*, 72(6):1208–1243, 2004.

[18] John Markoff. Attack of the zombie computers is growing threat. New York Times, January 7 2007.

[19] Ryan Naraine. EveryDNS under botnet DDoS attack. eWeek Security Watch, December 2006.

[20] Ryan Naraine. Is the botnet battle already lost? eWeek, October 2006.

[21] NETI@Home. `http://www.neti.gatech.edu/`, 2007.

[22] U.S. Department of Justice. *Searching and Seizing Computers and Obtaining Electronic Evidence in Criminal Investigations*. `http://www.cybercrime.gov/s&smanual2002.htm`, july edition, 2002.

[23] Phillip Porras and Vitaly Shmatikov. Large-scale collection and sanitization of network security data: Risks and challenges. New Security Paradigms Workshop, Schloss Dagstuhl, Germany, September 2006.

[24] PREDICT. `http://www.predict.org/`, 2007.

[25] 18 U.S.C. § 2707.

[26] Veyas Sekar, Yinglian Xie, David A. Maltz, Michael K. Reiter, and Hui Zhang. Toward a framework for internet forensic analysis. Proceedings of the 3rd Workshop on Hot Topics in Networks (HOTNETS-III), November 2004.

[27] Carl Shapiro. Exchange of cost information in oligopoly. *Review of Economic Studies*, 53:433–446, 1986.

[28] Adam Slagell and William Yurcik. Sharing computer network logs for security and privacy: A motivation for new methodologies of anonymization. SECOVAL: The Workshop on the Value of Security through Collaboration, 2005.

[29] Xavier Vives. Trade association disclosure rules, incentives to share information, and welfare. *RAND Journal of Economics*, 21(3):409–430, Autumn 1990.

[30] Tim Weber. Criminals may 'overwhelm the web'. http://news.bbc.co.uk/2/hi/business/6298641.stm, January 2007.

[31] Vinod Yegneswaran, Paul Barford, and Somesh Jha. Global intrusion detection in the DOMINO overlay system. In *Proceedings of Network and Distributed System Security Symposium (NDSS)*, February 2004.