Spamming for Science: Active Measurement in Web 2.0 Abuse Research

Andrew G. West[1], Pedram Hayati[2], Vidyasagar Potdar[2], and Insup Lee[1]

[1] Department of Computer and Information Science
University of Pennsylvania, Philadelphia, USA
{*westand, lee*}*@cis.upenn.edu*

[2] Anti-Spam Research Lab
Curtin University, Australia
{*p.hayati, v.potdar*}*@curtin.edu.au*

Abstract. Spam and other electronic abuses have long been a focus of computer security research. However, recent work in the domain has emphasized an *economic analysis* of these operations in the hope of understanding and disrupting the profit model of attackers. Such studies do not lend themselves to passive measurement techniques. Instead, researchers have become middle-men or active participants in spam behaviors; methodologies that lie at an interesting juncture of legal, ethical, and human subject (*e.g.,* IRB) guidelines.

In this work two such experiments serve as case studies: One testing a novel link spam model on Wikipedia and another using blackhat software to target blog comments and forums. Discussion concentrates on the experimental design process, especially as influenced by human-subject policy. Case studies are used to frame related work in the area, and scrutiny reveals the computer science community requires greater consistency in evaluating research of this nature.

1 Introduction

Spam needs little introduction given estimates that 95%+ of email traffic, 75% of all blog comments, and nearly every medium of human communication has been pervaded by the practice. The growing prevalence of distributed and collaborative models of information dissemination (*i.e.,* Web 2.0 forums, wikis, blogs, *etc.*) has only expanded the battleground. Measurement studies from the end-user perspective have long been the predominant method of examining these phenomena. More recently research has begun to consider the attacker's perspective: What are the motivations? How much money is made? What are the greatest marginal costs? By answering these questions researchers can hope to better understand the spam profit model and how to undermine it.

However, an empirical view of these notions does not come cheaply. The first-person viewpoints that enable such studies raise interesting legal, ethical, and human subject questions. Although formal bodies (*e.g.,* Institutional Review Boards, "IRBs") exist to regulate these matters it appears the computer science community is unfamiliar, or questioning of, their role. As two case studies reveal this yields unfair and inconsistent academic evaluations that satisfy neither authors, reviewers, or program committees (PCs).

This work begins by describing two recent works in this domain (Sec. 2), both actively conducting spamming campaigns with scientific interests. One targeted a collaborative platform (Wikipedia) and the other blog/forum environments. These case studies were subject to institutional review and their approval process is described at length (Sec. 3). This description: (1) sketches the approval process and policies that regulate this kind of research, and (2) outlines the experimental methodologies that brought these studies into compliance.

After the experiments were completed/described, the papers proceeded into the publication process. What followed exemplifies the inability of the community to soundly evaluate research of this kind (Sec. 4). Reactions ranged from applause to outrage; some endorsing IRB approval and others rejecting it entirely. It is unclear if the community is: (1) unfamiliar with the scope/role of the current review process, or (2) informed but dissatisfied with its judgments. Regardless, authors deserve a system by which research can be approved and evaluated under the same criteria. Similarly, reviewers are entitled to one that allows them to judge submissions on technical merit and not personal beliefs. This work continues by surveying literature about the evolving role of ethics, IRBs, and researchers in technical studies – and framing other active measurement work in this context (Sec. 5). Finally, concluding remarks are made (Sec. 6).

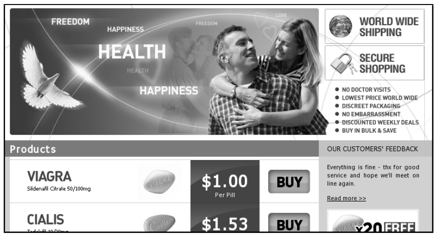
It should be emphasized that this work is a case study exemplifying ethical disagreement/issues. While it advocates the need for improvement, it does not endorse any particular mechanism for achieving it. While this remains an open issue for the computer science community, we believe it important to realize that the IRB is the only such regulator in the status quo.

2 Case Study Research

In this section we describe our case studies, two active measurement Web 2.0 link spam experiments which serve as the basis for later discussion. Included here is information about how these studies were conducted, the statistics collected, and the conclusions yielded by analyzing that data.

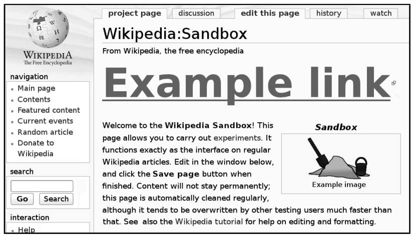
At a high level the experiments are quite similar with both economic components following the "pipeline" model described by Kanich *et al.* [13]. Summarily, after the spam hyperlinks have been disseminated there are three important measurements. First is the number of *exposures*, the quantity of individuals who view the spam link. Second is the *click-through* rate, the percentage of exposures that result in a visit to the *landing site* (*i.e.,* the webpage at the spam URL, or at the conclusion of that URL's redirection chain). Finally, the ratio of site visitors that actually make a purchase is termed the *conversion* rate.

Both case studies implemented a "payment disabled" store front (as per [13]) in order to collect the latter two statistics. These landing sites operate much like any online store but attempts to "check out" result in a technical failure or other



Everything is fine - thx for good service and hope we'll meet on line again.

Fig. 1. Example landing site



Wikipedia:Sandbox

From Wikipedia, the free encyclopedia

Welcome to the **Wikipedia Sandbox**! This page allows you to carry out experiments. It functions exactly as the interface on regular Wikipedia articles. Edit in the window below, and click the **Save page** button when finished. Content will not stay permanently; this page is automatically cleaned regularly, although it tends to be overwritten by other testing users much faster than that. See also the Wikipedia tutorial for help on editing and formatting.

Fig. 2. Prominent link display in a *wiki*

complication. In this manner, one can approximate purchase quantity and value without having to fulfill transactions. Both case studies constructed/scraped landing sites that were pharmaceutical in nature (see Fig. 1).

The two experiments[3] differ most in the environments being studied. The first, conducted at the University of Pennsylvania, performed proof-of-concept attacks to show the economic viability of a *novel* spam model against the collaborative encyclopedia, Wikipedia (Sec. 2.1). The second, via Curtin University, used blackhat software to target web forums and blog comments (Sec. 2.2).

2.1 UPenn Research: Wikipedia Spam

Collaborative functionality is becoming increasingly prevalent in web applications and no paradigm embodies this more purely than the *wiki*. The open-editing permissions and massive traffic[4] of some wiki installations (*e.g.,* English Wikipedia) would seem to invite spam behaviors. To the surprise of its authors, a measurement study [25] found status quo spam behaviors to be technically naïve, comparatively infrequent, and ineffective for their perpetrators.

Using their expertise of collaborative security the researchers next sought to identify vulnerabilities of the platform. In doing so they described a novel attack model that exploits the latency of Wikipedia's human-driven enforcement [25]. In this model link placement is characterized by: (1) targeting high traffic pages, (2) prominent link styling (see Fig. 2), and (3) the use of privileged accounts.

To show the viability of the attack model an active measurement study was engineered. The experiments added links to a payment-disabled pharmacy per the proposed strategy. Only seeking to establish a proof-of-concept, just 3 Wikipedia accounts were used. These accounts posted 341 hyperlinks with each surviving for an average of 93 seconds. Public article view statistics show that ≈14,000 individuals were exposed to the link, generating 6,307 click-throughs (*i.e.,* landing site visits) that led to 8 "purchases" for $1940 USD.

The "revenue" generated considerably exceeded the marginal attack costs, suggesting a viable attack model (at least initially). Despite IRB approval, these

[3] The authors of *this* paper are a subset of those conducting the case study research.

[4] Distributed attacks that target low traffic and poorly maintained wikis for search-engine-optimization (SEO) are not uncommon. The research under discussion, however, concentrates only on direct traffic (*i.e.,* click-throughs) in high exposure wikis.

results remain unpublished for reasons discussed in Sec. 4. However, this result did motivate additional research into protections against such vulnerabilities, the suggestions of which have been actively described and implemented [24].

2.2 Curtin Research: Blog/Forum Spam

Relative to the novel proposal on Wikipedia, link spam in blog comments and forums is a pervasive issue. One source estimates that 75%+ of blog comments are spam [4]. This proliferation suggests the status quo attack model is profitable to its perpetrators, motivating research into its economic dynamics. We choose to highlight [10, 11], which actively posted links to these environments.

Work began by *harvesting* sites running common software (*e.g.,* phpBB, Wordpress) with circumvent-able protections (*e.g.,* CPU-solvable CAPTCHAs, no registration, *etc.*). In this manner the common structure and weaknesses can enable autonomous link placement at minimal marginal cost. Such functionality has been encoded into blackhat software and the experiment used one such tool: XRumer [3, 21]. In addition to posting links the researchers also spoofed the "referrer URL" of the HTTP headers to point to the pharmacy site[5].

The harvesting stage produced a list of ≈98,000 websites of which 66,226 were practically targeted after discounting network errors. From these, 7,772 links (11.7%) were successfully posted to public view, with the remainder being caught by spam filters, manual moderation, *etc.* The month-long experiment produced 2,059 pharmacy visits and 3 "intents to purchase." Minor modifications in link placement strategy also permitted more fine-grained observations. For example, non-English websites produced higher response rates, and referrer spam produced more landing site hits than the actual link placements.

As of this writing the recent research remains unpublished. However, it is the authors' intention for the work to appear as [10].

3 Obtaining Research Approval

Our case studies now summarized, we next describe their formal approval process. More than rote description, this discussion intends to use the approval criteria as an outline for focusing on experimental design and ethical issues. We begin by justifying the need for active measurement (Sec. 3.1). Having decided to use human subjects, we next describe the approval workflow (Sec. 3.2). Then, we handle the talking points of approval: informed consent (Sec. 3.3), maintenance of privacy (Sec. 3.4), and minimization/justification of harm (Sec. 3.5).

3.1 Infeasibility of Passive Measurement

Before engaging in active measurement it should be the case that a passive approach is not feasible. A leading study of email spam economics [13] creatively

[5] This is a technique called "referrer spam", "log spam", or "referrer bombing." Sites that make access logs public will have the spam URLs indexed by search engines.

became a “man-in-the-middle” to a botnet operation and rewrote spam URLs to a payment-disabled pharmacy under their own control. In this manner, no additional spam was sent and the spam they rewrote was less malicious than it would have been otherwise. A similar strategy is difficult to imagine in Web 2.0 environments where attacks are coordinated by software empowered individuals.

Recruitment of cooperative blog/forum owners for research purposes deserves consideration. No additional spam would need to be injected as status quo events could be examined. Visitor logs would quantify exposure and outgoing link clicks could be tracked. However, this presents issues: (1) participating owners are unlikely to form a representative set (poorly maintained sites are likely crucial for attackers), and (2) this result says nothing about conversion rates.

The Wikipedia study has additional complications. Given a single intended target (English Wikipedia), a rejected request for cooperation would raise administrative awareness and bias any subsequent (non-consenting) trials. Moreover, because the strategy is a novel one it is impossible to glean statistics without injecting links per the proposed model.

3.2 Approval Workflow

Having decided to undertake active measurement, formal approvals must be obtained from organizations overseeing: (1) human-subjects/ethics and (2) legality.

Human-subjects/ethics: Any experiment involving data collection from humans is required to undergo review. Internationally these groups go by different names but are quite similar in function; the U.S. has the Institutional Review Board (IRB), Australia prefers Human Research Ethics Committee (HREC), and the European Union uses Research Ethics Committees (RECs).

There is ongoing debate over whether human subjects approval is equivalent to an experiment being ethical or whether it is only a component thereof. We challenge readers to imagine any form of ethically interesting research that does not at least indirectly impact humans (or animals) in some way (physically, psychologically, economically, *etc.*). Regardless, some in the computer science community do draw this distinction leading to inconsistency in the evaluation of research (see Sec. 4). Further examination of this controversial issue is beyond the scope of this work, as we prefer to focus on experience-driven analysis.

In the blog/forum case study (Curtin University, AUS), the process began by contacting a department-level ethics coordinator. This individual determined the experiment to be “low risk” and eligible for an expedited review. Per University/AUS policy [1] low risk research is that which “does not pose a greater risk than participants would face in their normal daily routine.” Supporting this criterion are the facts that: (1) advertisements and spam are already ubiquitous in blog/forum environments, and (2) statistical collection on the web is omnipresent. After one meeting the study was allowed to continue.

Matters were more complex for the Wikipedia case study (UPenn, USA). After an IRB coordinator found that the research posed “more than minimal risk to subjects” [2] a request for expedited approval was rejected in favor of a full board review. Seemingly, the concern was that publication of the novel attack model could considerably endanger Wikipedia's operation if the vulnerabilities remained unpatched (see Sec. 3.3). After multiple iterations of clarification and research gathering the protocol was approved in ≈14 weeks time[6].

This "full board review" produced a number of observations which may be interesting to readers. First, the board proceedings are non-transparent and closed-door (except for clarifications), doing little to inform other researchers how to best shape their experiments to the satisfaction of the IRB/HREC/*etc.* Further, the latency and lack of technical expertise among members have previously been identified as weaknesses of the process [6, 9].

Legal Approval: The legal approval process was less structured. The Wikipedia study did come to the attention of the University's Office of the General Counsel, who did not object to publication of the study results with IRB approval. The blog/forum research was not required to seek such approval by their coordinator. The legal framework in which this research operates is beyond the scope of this work (see [7]), though it is interesting to consider how differing jurisdictions may affect what is deemed "acceptable" research.

3.3 Regarding Participant Consent

A majority of human subjects studies operate under *informed consent*, whereby a potential subject is informed a priori of the purpose and potential risks of participation. If he/she voluntarily decides to proceed, this removes considerable responsibility from the researcher. It is possible to forego informed consent where it is: (1) technically impractical and/or (2) biasing of results. However, this places stricter requirements on the experimental methodology. Both case studies operated without the prior consent of any participant.

As discussed in Sec. 3.1, contacting site *administrators* would create recruitment bias and/or raise administrative awareness. In the case of *readers*, informed consent also produces numerous issues. Consider that experiments take place on a 3rd-party site where: (1) the consent dialogue alone might constitute spam, and (2) limited control would force that dialogue to be awkwardly adjacent to the behavior being measured. Further, those who choose to ignore the spam messages (the vast majority of exposures) incur minimal disruption. One might imagine that forcing everyone to opt-in/out of the experiments would create more annoyance than the experiments themselves.

Following these arguments, both case studies were approved to proceed without informed consent. Having chosen this course, the anonymity of the exposures/readers becomes paramount (Sec. 3.4). This does not eliminate the pos-

[6] Experiment design was influenced by the ethical norms of the IRB process. However, it should be acknowledged that approval was received in an ex post facto fashion, due in part to an initial miscommunication with the IRB. Such ex post facto scrutiny follows the same workflow and is held to the same standard as a priori review. This occurrence speaks to the unfamiliarity and poor working relationship others have reported between computer scientists and these boards (see Secs. 4 and 5).

sibility of *debriefing* test subjects after their participation. Readers could potentially be notified as they exit the experiment pipeline (navigating off-site; attempting to purchase) but such information could influence how others interact with the experiment. For example, in a wiki setting, a reader who discovers the "spam" to be an academic experiment may not give it treatment consistent with spam links (*i.e.,* removal). Notification en masse after the entire experiment duration is not possible given the decision to preserve anonymity.

In the case studies, one instance of debriefing was present. The administrative community of Wikipedia (the Wikimedia Foundation) was contacted post-experiment. In this email notification the vulnerabilities were described and technical assistance was offered towards mitigating the exploit.

3.4 Privacy and Data Security

Given that we have collected the behavior of non-consenting users, there is a responsibility to protect that data: its release could lead to embarrassment or other harm. One way to prevent this is by severing the mapping between experiment events and real persons (*i.e.,* an anonymous experiment).

In the Wikipedia experiment under the IRB system, information capable of identifying real persons is called *personally identifiable information* (PII). The IRB required that data collected by the checkout system be immediately destroyed (it never left the client machines) with the exception of the items being purchased and their value. To uniquely identify click-through and purchasing users, a hash of the IP address was stored[7]. The IP addresses themselves were considered PII due to static IPs, the possibility of geo-location, *etc.*

The Australian notion of privacy had a very different interpretation. In the blog/forum case study server logs were maintained and authors geo-located their landing site visitors. In their setup, registration was required to checkout, with relevant fields including: (1) first name, (2) last name, and (3) email address. These fields were manually *inspected* to ensure the registration attempts were legitimate, before the data was destroyed. The Australian body seems to operate on the logic that "since normal spam sites would view registration data, it is permissible for the researchers to do so."

Just as participant data must be secured there is a need to protect the identities of the researchers and their institutions mid-experiment. Thus, we discuss the computing framework in which these studies operated. Consider that it is desirable to use non-institutional IP addresses to launch the experiments and host the landing site. This avoids experimental bias (*e.g.,* `*.edu` sites might not trip filters) and protects the institution from ill consequences (*e.g.,* blacklisting). One case study launched experiments from a large cloud provider and hosted their landing site via a 3rd party service (whose data retention was vetted). The other study purchased a dedicated Internet connection (outside the University network) for hosting and used proxy servers and VPN for outbound traffic.

[7] As one reviewer pointed out, the finite nature of IP space makes it feasible to reverse these (now destroyed) hashes – a consideration not foreseen in experiment design.

3.5 Minimizing and Justifying Harm

Harm in any experiment should be both *minimized* and *justified.* We now extend previous discussion about risk minimization to include experiment elements that were not major "talking points" of the approvals process.

Experiment Scale: Rather than assessing risk at the per-subject level, a more pragmatic approach is to consider the cumulative cost to all participants, making the *scale* of experiments a significant factor. The blog/forum case study was approved without any conditions on the size of the experiments (though ethical approval is valid for 12 months, after which re-evaluation is required). In those experiments 66,000 sites were targeted, a scope seemingly justified by the large number of sub-experiments and need for statistically significant data. Consider that while the number of targets/exposures is large, relatively few engage in the interesting behaviors (click-through, conversion) being measured.

Generally one should carefully weigh the need for statistical significance against human costs. Showing the viability of novel theories should require less iterations than measurement studies (although the Wikipedia study produced 14,000+ exposures in just 3 trials). Also consider that long running or narrowly focused experiments could target the same individual(s) multiple times.

Deceptive Advertising: Ethical review boards tend to be sensitive to *deception* of test subjects. At the same time, attackers are by their very nature deceptive agents and accurate simulations need to reflect this nature. In the case studies one potential source of deception is hyperlink presentation, *i.e.,* the *hooks* or description that is associated with links. Among several strategies it was the alluring and deceptive hooks (*e.g.,* "click to collect your prize") which proved most controversial. However, hooks of this type far outperformed more mundane approaches, speaking to the effectiveness of such social engineering tactics.

Others might question the choice of landing site genre. One case study sold a wide range of pharmaceuticals while the other focused only on the "male enhancement" subset. A previous study [14] showed it is precisely these products which dominate online spam revenue. Moreover, care was taken to make sure the sites were free of any harmful/suggestive imagery and descriptions.

Just because harm is minimized (while maintaining experiment integrity) does not mean it is *justified.* For that to be true, experiments must produce a net benefit which exceeds any harm (a *consequentialist* approach [8]). This is a particularly unsatisfying condition given that the outcomes of the research cannot be known a priori. Nonetheless, in the case studies the: (1) novelty of the work, (2) daily exposure of readers/administrators to spam behaviors, and (3) projected understanding of the spam ecosystem indicated a foreseeable benefit.

Project benefit can be more accurately assessed in an ex post facto fashion. Wikipedia active measurement showed the attack model viable, motivating the authors to create a spam detection engine for wikis [24]. A live implementation of the technique has already assisted in the removal of far more spam instances than placed during active measurement. Moreover, circumstances arising during the experiment encouraged further study into "redacted revisions" [26].

4 Community Response and Discussion

Once case study research was completed the logical next step was to submit the findings to academic journals and conferences. The inconsistent treatment of ethical issues in reviewer feedback was not expected. More importantly, it raises questions about how the community and publication process can better accommodate research of this kind.

Reviewer Response: The response to Wikipedia active measurement was extremely mixed[8] (all submissions noted the IRB approval). Some reviewers applauded the study, finding the methodology appropriate and necessary proof of an earlier hypothesis. Others took a more neutral approach, pointing to possible ethical implications but stating that the IRB approval was evidence of reasonable conduct. Others still assaulted the methodology, questioned the social conscience of the authors, and were prepared to reject the paper on ethical grounds alone. Excerpts from some of the critical responses are in Appendix A.

Several iterations of submission followed. In one attempt, several pages were dedicated to ethical justifications (pages that could have been dedicated to technical content). In the end, it was decided to omit the active measurement results from the paper due to these complications and concerns over statistical significance/stability. In the published version [25] there is only a numerical estimation of attack viability. Though only just beginning the publication process the blog/forum study is experiencing similar reactions.

Discussion: A major issue is why some reviewers are not satisfied with ethical review decisions and make it their own responsibility to regulate the matter. An IRB is best-equipped to make these decisions, being armed with experience and precedent, and having seen supporting documents to which reviewers are not privy[9]. This may be partially explainable by the unfamiliarity many computer scientists have of these organizations, as described by Garfinkel [9].

One such example can be seen in [5] where the author suggests IRBs are insufficient and bases this on a flawed argument. He cites two experiments that "do not involve human subjects...": (1) an experiment that congested residential Internet networks to learn about their characteristics and (2) a study that de-anonymized packet traces, linking them to physical addresses. We believe strongly these are human subjects issues that fall under IRB/committee jurisdiction. The first example would affect Internet QoS for users and the second has obvious privacy implications.

We acknowledge that the IRB (and its equivalents) may be a *logistically* imperfect system (see [9]). However, in the absence of an alternative, this is no

[8] It is difficult to quantify the weight these ethical disagreements had in accept/reject decisions (although one reviewer did make the connection explicit, see Appendix A). We prefer to focus solely on the qualitative feedback given about active measurement.

[9] Submitted versions included a footnote indicating that reviewers/PC-members could be contacted to obtain a copy of the approval documents (*e.g.,* via the conference chair to preserve anonymity). No such requests were made.

reason not to respect its findings. Allowing PC chairs or reviewers to interject their beliefs only lends greater subjectivity to evaluation. Consider that research similar to the case studies has been published in spite of complaints and *without* IRB approval (Sec. 5.1). Such a situation is unimaginable in many fields, where IRB approval is considered a gold standard of approval (see again, Sec. 5.1).

For those who advocate a more responsive and technically-staffed IRB-like organization it is clear it should come into force *before* research is conducted. The current situation creates awkward situations where research has been performed (along with any harm) but cannot be released to the community. Such an organization also faces practical challenges in creating an objective and level playing field. For example, how does one integrate the legal frameworks of international researchers? Will the organization supersede or operate alongside the human subject boards (adding bureaucracy)? Who writes the policies?

Such an organization could be an asset to the community but is far from being realized. Focusing on the status quo, human-subjects boards are the only organizations properly equipped to handle these matters (and arguably, already do so at the appropriate scope). This division-of-labor allows reviewers/PCs to concentrate on their area of expertise: technical merit. Although imperfect, these boards are the most satisfactory regulators of research ethics at this time. As such, respect for their decisions is the greatest hope the community has for fairly evaluating ethically interesting research.

5 Related Work

Discussion of related literature begins by looking at other spam and electronic abuse research that has employed ethically notable active measurement techniques (Sec. 5.1). Then, we look at writings about the formal review process and issues specific to computer science research (Sec. 5.2).

5.1 Similar Abuse Research

As other active measurement studies are surveyed, we encourage readers to think about experimental design and the potential risk posed. We divide our review into: (1) studies that have engaged in spam-like behaviors, and (2) studies that involve payment to spammers or spam-support services. To permit discussion our literature selection is both non-exhaustive and narrow. Readers are encouraged to see the survey of Moore and Anderson [18] for a broader look at ethically-interesting security research, particularly of the empirical and behavioral variety.

Studies Conducting Abuse: The most similar work to the wiki case study is [22] wherein the authors befriended 942 popular individuals on a social networking site. Then, they posted a “comment” including a 1×1 pixel image hot-linked from their own server (the image(s) were sometimes 50+MB in size). In their 12-day experiment their server recorded 2,598,692 hits, indicating the feasibility of conducting DDOS attacks in this fashion. Though readers’ attention was not particularly affected (their bandwidth was), administrator workload was non-trivial. Communication with the authors revealed the work did *not* have IRB approval and some reviewers raised complaints, yet the paper was still published.

Another work looked at the topic of "social phishing" [20]. There, the researchers mined social network data in order to write personalized phishing emails, sent to 600 university students. Relative to a control these customized emails produced higher "success" rates (with 72% of students providing their university credentials). This research had IRB/institutional approval as the paper discusses at length, along with test-subject reactions to the work.

An interesting cross-domain perspective comes from [17] where the authors posed as prospective students, emailing 6,600 university professors requesting a meeting. Student names were chosen to imply gender/race, with the research measuring the varying response rates. While not commercial spam, one could imagine the cost per participant was quite high (reading the email, responding, meeting scheduling/cancellation). This study had the IRB approval of multiple institutions but drew considerable criticism in Internet communities. In an interesting contrast to the case studies much condemnation was directed not at the researchers, but the IRBs involved. Similarly, communication with an author indicated no reviewer had raised ethical complaints.

Finally, [13] deserves mention for inspiring much research on active measurement of electronic abuse. Therein the authors became a coordinating node in a spam botnet. From this position they instruct worker nodes to send the same spam emails they would have otherwise, but change the hyperlink URL to one under their control (a payment-disabled pharmacy). However, because the study "strictly reduced harm" (no new spam; made sent spam less malicious) it lies on less tenuous ethical footing than the other work described herein.

Studies Aiding Abusers: Another frame of reference into spam economics can be achieved by becoming a consumer of spam services. Just like one of the case studies, [21] purchased blackhat spamming software (at $400+ USD) in order to analyze its operation. Another work [19] spent hundreds of dollars to solve 100,000+ CAPTCHAs to study the dynamics of that underground economy. Finally, in [14, 15] researchers made 156 purchases from spam-advertised business to make inferences about sale volume and examine financial routing. It is especially hard to quantify the harm that may be indirectly suffered as a result of financially assisting these individuals/services. However, as evidenced by the above papers, this seems to be a tactic generally well-received by the community.

5.2 IRB/Ethical Discussions

Numerous works have looked at legal, ethical, and human subjects issues in computer science research. None is more relevant than [8], which lends a broader perspective to the experiences shared herein. That work identified the weaknesses/limitations of the status quo to be: (1) an absence of shared community values, (2) lack of familiarity with ethics and review systems, and (3) lack of consensus on enforcement. They too looked at ways the community could move forward, suggesting self-governance, public discussion, and protocols to reward ethical behavior. Outside the scope of our writing, [8] also considers the roles of professional societies (*e.g.,* ACM, IEEE) and funding organizations.

Other writings have more narrow scope. Focusing on legal issues in particular is [7], emphasizing the collection/sharing of network traces. The IRB has been a point of emphasis, beginning with a look at how the Internet has changed its role [23]. Other works [6, 9] criticize the IRBs latency and lack of technical expertise, with the latter claiming that much CS research runs afoul of regulation. Moving beyond the IRB, [5] examines the program committees role in ethical evaluation. Then, there are "best practices" papers like [16], focusing on "vulnerability research". Finally, Kanich [12] writes similarly based on his extensive experience with economic and cyber-crime research.

6 Conclusions

In this work, two case studies guided a discussion of the legal, ethical, and human-subject considerations of active measurement research in spam and electronic abuse. Much discussion was dedicated to how experimental design was shaped by the review process, bringing the controversial methodologies into policy compliance (of the IRB or its international equivalents). We intended this to give some introduction into the role/operation of these review committees and inspire readers to think about increasingly benign ways to gather data.

Paper rejections, negative reviews, and harsh personal criticism are not something many authors are eager to speak about. However, in relaying our own experiences we hope to give exposure to an issue on which the computer science community can improve: evaluating ethically interesting research. Critics condemn the IRBs latency, handling of technical matters, and scope. If this is indeed a widely held view the review stage is a poor place to enforce it, and new bodies need to be assembled to proactively regulate these matters. If no such consensus exists (or until such a body is in place) then the community should respect the current standard. Either way, the status quo is detrimental to authors, reviewers, PCs, and the entire community – and the exposure and elimination of this practical dilemma was our motivating interest in authoring this work.

Acknowledgments. The authors would like to thank their colleagues who contributed to the original case study research: Jian Chang, Krishna Venkatasubramanian, and Oleg Sokolsky from UPenn [25]; Nazanin Firoozeh and Kevin Chai at Curtin [10]. The authors also acknowledge the helpful advice and service of their respective human-subject/ethical/legal coordinators and boards. In particular, Robert R. Terrell of UPenn's Office of the General Counsel is thanked for his guidance. Any opinions expressed in this work do not necessarily reflect the sentiments of those acknowledged here.

References

1. Curtin: Research management. http://research.curtin.edu.au/guides/
2. UPenn: Office of regulatory affairs. http://www.upenn.edu/regulatoryaffairs/
3. XRumer. http://www.xrumerseo.com/, (Blackhat SEO software)
4. Abu-Nimeh, S., Chen, T.: Proliferation and detection of blog spam. IEEE Security and Privacy 8(5), 42–47 (2010)
5. Allman, M.: What ought a program committee to do? In: USENIX Workshop on Organizing Workshops, Conferences, and Symposia for Computer Systems (2008)
6. Buchanan, E.A., Ess, C.M.: Internet research ethics and institutional review boards: Current practices and issues. SIGCAS Computers and Society 39(3) (2009)
7. Burstein, A.J.: Conducting cybersecurity research legally and ethically. In: LEET: Proc. of the Wkshp. on Large-Scale Exploits and Emergent Threats (2008)
8. Dittrich, D., Bailey, M., Dietrich, S.: Building an active computer security ethics community. IEEE Security and Privacy 9(4) (July/August 2011)
9. Garfinkel, S.L., Cranor, L.F.: Institutional review boards and your research. Communications of the ACM 53(6), 38–40 (June 2010)
10. Hayati, P., Firoozeh, N., Potdar, V., Chai, K.: How much money do spammers make from your website?, (Working paper, in submission)
11. Head, B.: Storage bills top $43,000 say spam-busters. ITWire.com (August 2011), http://www.itwire.com/business-it-news/security/49239-storage-bills-top-43000-say-spam-busters
12. Kanich, C., Chachra, N., McCoy, D., Grier, C., Wang, D., Motoyama, M., Levchenko, K., Savage, S., Voelker, G.M.: No plan survives contact: Experience with cybercrime measurement. In: CSET '11: Proceedings of the 3rd Workshop on Cyber Security Experimentation and Test (August 2011)
13. Kanich, C., Kreibich, C., Levchenko, K., Enright, B., Voelker, G.M., Paxson, V., Savage, S.: Spamalytics: An empirical market analysis of spam marketing conversion. In: CCS'08: Proc. of the Conf. on Computer and Comm. Security (2008)
14. Kanich, C., Weaver, N., McCoy, D., Halvorson, T., Kreibich, C., Levchenko, K., Paxson, V., Voelker, G.M., Savage, S.: Show me the money: Characterizing spam-advertised revenue. In: Proc. of the USENIX Security Symposium (August 2011)
15. Levchenko, K., Chachra, N., Enright, B., Felegyhazi, M., Grier, C., Halvorson, T., Kanich, C., Kreibich, C., Liu, H., McCoy, D., Pitsillidis, A., Weaver, N., Paxson, V., Voelker, G.M., Savage, S.: Click trajectories: End-to-end analysis of the spam value chain. In: Proc. of the IEEE Symposium on Security and Privacy (2011)
16. Matwyshyn, A.M., Cui, A., Keromytis, A.D., Stolfo, S.J.: Ethics in security vulnerability research. IEEE Security and Privacy 8, 67–72 (2010)
17. Milkman, K.L., Akinola, M., Chugh, D.: The temporal discrimination effect: An audit study of university professors, (Working paper)
18. Moore, T., Anderson, R.: Economics and Internet security: A survey of recent analytical, empirical and behavioral research. Tech. Rep. TR-03-11, Harvard University, Department of Computer Science (2011)
19. Motoyama, M., Levchenko, K., Kanich, C., McCoy, D., Voekler, G.M., Savage, S.: Re: CAPTCHAs - Understanding CAPTCHA-solving services in an economic context. In: USENIX Security Symposium (August 2010)
20. Nathaniel, T.J., Johnson, N., Jakobsson, M.: Social phishing. Communications of the ACM 50(10) (October 2007)
21. Shin, Y., Gupta, M., Myers, S.: The nuts and bolts of a forum spam automator. In: LEET: Proc. of the Wkshp. on Large-Scale Exploits and Emergent Threats (2011)

22. Ur, B.E., Ganapathy, V.: Evaluating attack amplification in online social networks. In: W2SP'09: The Workshop on Web 2.0 Security and Privacy (2009)
23. Walther, J.B.: Research ethics in Internet-enabled research: Human subjects issues and methodological myopia. Ethics and Info. Technology 4(3), 205–216 (2002)
24. West, A.G., Agrawal, A., Baker, P., Exline, B., Lee, I.: Autonomous link spam detection in purely collaborative environments. In: WikiSym '11: Proc. of the Seventh International Symposium on Wikis and Open Collaboration (October 2011)
25. West, A.G., Chang, J., Venkatasubramanian, K., Sokolsky, O., Lee, I.: Link spamming Wikipedia for profit. In: CEAS '11: Proc. of the Eighth Annual Collaboration, Electronic Messaging, Anti-Abuse, and Spam Conference (September 2011)
26. West, A.G., Lee, I.: What Wikipedia deletes: Characterizing dangerous collaborative content. In: WikiSym '11: Proc. of the Seventh International Symposium on Wikis and Open Collaboration (October 2011)

Appendix A: Reviewer Comments

Below is a sample of reviews received in response to the Wikipedia line of link spam research. Effort has been made to preserve the context of the feedback. Each bullet point represents the comments of a single reviewer.

- "The second measurement study is a bit offensive, but the IRB approval seems to cover this ... While the IRB problem is discussed, I am still not convinced that such experiments with Wikipedia are good from an ethical point of view."
- "I personally am concerned about the ethics of the active link-spamming research ... In particular, a natural guideline is that research should not cause harm or damage to subjects without their informed consent. In this study, it appears that harm or damage may have been done to Wikipedia by this research ... and was done without prior consent of the Wikipedia foundation ... not persuaded that the 'consequentialist' viewpoint is a suitable response to this concern."
- "Although they did get their institution's IRB to approve it, IRB approval is a necessary, but not sufficient, step for justifying such an experiment ... the experiment imposed a substantial cost on the Wikipedia community, both the editors who had to fix the page, and the thousands of users who encountered their spam. Such a cost, which is involuntary to the participants, needs to be justified by a significant gain in scientific understanding."
- "... their active experiment is ethically deficient ... I view each [of multiple issues, the 'ethical deficiency' included] as a deal-breaker ... The ethical standing is dubious enough that it does *not* suffice to simply tell us that you had IRB approval. We need to know the wording of what the IRB approved. In addition, while the text briefly mentions (un)informed consent, there is no mention of *post facto debriefing* ... [this] makes the reviewer wonder to what degree the authors really did obtain IRB approval that was itself informed."
- "... the paper is rather offensive, it seems like Wikipedia actually received negative press related to this experiment ... I find this a bit questionable, the discussion in the appendix is also not very convincing. Actually I had not thought that the authors would receive IRC approval for this kind of study. I suggest to revise the appendix and maybe even publish all IRC documents ... Apart from this aspect, the study is interesting and the authors demonstrate convincingly that Wikipedia is an attractive target for link spam."