

## Introduction

From the standpoint of an industrious criminal, cybercrime is a particularly attractive route to illegitimate wealth. We all recognize the typical Nigerian-Prince-has-bequeathed-you-a-fortune ploys, but do ourselves a great disfavor by identifying these as the sole or primary vestige of malicious internet enterprise. However, dealing in fear-mongering over *Wargames*-style insecurity of nuclear security infrastructure is also a grossly misleading characterization. Some scholars have begun to investigate the technical details and incentives for state and non-state actors to engage in acts of cyberwarfare (Carr and Goel, 2010). Others have made qualitative analyses of small-time cyber criminal enterprise (C´ardenas et al. 2010). Few, if any, have investigated the smaller-scale cybercrime in formal models. Specifically, this paper uses these tools to investigate the very common problem of spam.

Spam is much more akin to the aforementioned Nigerian-Prince scam than organized cyberwarfare, though the tools for committing either are common to both. For the purposes of this paper, I use *spam* to denote any instance of an individual sending out massive amounts of text which is intended to be visible to other people, either via e-mail, SMS, chats, forum posts, comments, or wiki defacement. A *spammer* is simply one who either engages in this activity, or may do so.

The criminals of the internet are rational, smart, and motivated singularly by the drive to turn your money into their money. The architectures behind the scams supporting these operations are varied, as are the venues through which they are seeded, but the fundamental efforts of spammers are the same: reach as wide an audience as possible. Because of these generalizations, spam can be modeled as a simple strategic game.

### Modeling Cybercriminal Behavior

At the most fundamental level of spam, there are two actors: the spammer and the potential victim. The spammer has two options: illegitimately solicit his/her scams (spam) or not solicit. The could-be recipient can respond, or not. Substantively, this set of actions may seem inherently nonsensical at first. Implicit within the game is the possibility that the *victim* could seek out the *spammer* (a la the (Not Solicit, Respond) action profile). While this prospect seems foolish, there are situations to explain it. Some people, enticed by the potential to get rich quickly, actually will seek out spammer-run scams looking for the same opportunities as those who were simply caught by curiosity after having the spam e-mail, comment, or instant message delivered to them. The payoffs are modeled in the matrix below.

	Respond	Not Respond
Solicit	g, -g	0, 0
Not Solicit	g, -g	0, 0

In this simple model, g is monetary gain. There are two Nash Equilibria: (Solicit, Not Respond) and (Not Solicit, Not Respond). In other words, both situations in which the potential victim declines to respond. However, without anyone with an incentive to respond, the model describes a world in which spam may or may not exist, but spammers have neither the incentive to spam nor the incentive not to. Thus, it is likely more appropriate to ignore the very small class of individuals who will unwittingly seek out scams, and thus revise the model.

	Respond	Not Respond
Solicit	g, -g	0, 0
Not Solicit	0, 0	0, 0

Boyles

Now we have a single equilibrium which more closely resembles reality: Spammers have no incentive to not solicit, and the recipients have no incentive to respond, yielding (Solicit, Not Respond). This is exactly what happens in the vast majority of cases. It is, however, still imperfect. The cases in which a victim seeks out a scam are hugely marginal, but they do still exist. It would be more appropriate to discount them from this model and evaluate this interaction as an extensive game with imperfect information, in which the spammer's decision not to spam leads directly to the outcome (0,0). That said, an imperfect model is not synonymous with a useless one.

There are a few assumptions within the model; specifically, the cost of *sending* spam is practically trivial, enforcement of anti-spam law is nearly impossible, and potential payoffs are enormous. If we assign probability 'p' to the probability of solicitation, and 'o' to the probability of response, we can generate an expected value function for the spammer's actions:

$$E(\text{Solicit}) = (o)(g) + (1-o)(0) = og$$

$$E(\text{Not Solicit}) = 0$$

So the spammer has no reason *not* to spam, so long as g is positive (which will always be the case). And even though the expected return is very small (o is a shrinking probability as more people become aware of more spam methods and use more aggressive spam filters), one spammer repeating the game hundreds of thousands of times will readily generate a handsome sum.

This model has thus far examined spam as a zero-sum game. Reality is not so tidy. While the cost of sending out spam is effectively trivial, the cost to the spammer of the wrong person receiving it can complicate the game significantly.

### Unforeseen Consequences

A typical spammer runs a *botnet* (a network of computers with compromised security, called bots, with resources devoted to generating and spreading spam) to reach as wide an audience as possible. However, no botnet is yet clever enough to filter out security experts. As a byproduct, security experts can provide vigilante countermeasures. Even a cursory analysis includes a *WhoIs* report, which generally lists the personal details of the registrant. In other words, a lazy or novice spammer may leave enormous amounts of personal information at the disposal of someone with the knowledge and desire to find it. This information can be turned over to law enforcement agencies, or used to manually mount counterattacks to bring down the botnet infrastructure. At the most cursory level, the spammer's information can be posted to the wider internet at large, making him/her a very attractive target for other spammers and cybercriminals.

The drive to retaliate is a difficult thing to measure. In purely financial terms, it can be estimated to be approximately equivalent to the following matrix.

	Retaliate	Not Retaliate
Solicit	-d, {0,-c}	0, -c
Not Solicit	-d, 0	0, 0

In this payoff matrix,  $d$  is the cost of the damage a motivated security expert can cause and  $c$  is the technical cost of permitting the solicitation. The value of  $c$  is generally small—little more than some annoyance, moderate security risk, or increased bandwidth costs. Note that retaliation may yield either the same cost to a security expert as not doing so, or it may result in the spammer relenting for a net change of 0.

Boyles

This construction overlooks one important detail about the mentality of the expert: unlike the spammer, money is not his/her primary currency. The security expert is much more concerned with protecting his own network resources (and in many cases, the pride of keeping the network secure), which involves two steps: securing his own network against further attacks of the same or a similar nature, and debilitating the spammer, if possible. The most obvious route to doing the latter is giving the details of the incident to the appropriate law enforcement agency, but most effective spammers operate in regions where law enforcement cannot cope or does not care about digital theft. China, Southeast Asia, North Africa, Eastern Europe and the Middle East are all havens for spammers. If law enforcement cannot be compelled to intercede, it falls back to the expert to fend for him/herself. Elaborate technical counterattacks or simply “blowing the identity” of the spammer may yield a drop-off in the attacks, or an open communication channel for bargaining.

These values would seem to clearly suggest the best strategy is to simply avoid the internet underground altogether. However, this is like the original game in that it creates a rare and unlikely class of individuals: those security experts who seek out spammers unsolicited. While such people do exist, their number is far too small to play a meaningful role in the decision calculus of the spammer. Thus, it is more realistic to change the payoffs of (Not Solicit, Retaliate) to (0, 0). This warps the game into a world where spammers never spam for fear of retribution at the hands of the security expert, who will always retaliate if provoked, and always do nothing if not.

It is important to keep in mind that this game is played out between a *security expert* and a *spammer*. Of any thousands of potential victims, only a handful are likely to have the strong background in digital security necessary to mount a meaningful retaliation. Likewise, only a

Boyles

handful are naïve enough to respond to the spam. If we generalize this to a game intended to be played between the spammer and each potential victim (security mind or not), the game begins to become much more interesting.

	(o) Respond	(q) Retaliate	(1-q-o) Neither Respond nor Retaliate
(p) Solicit	$g, -g$	$-d, \{-c, 0\}$	$0, -c$
(1-p) Not Solicit	$0, 0$	$0, 0$	$0, 0$

(Note, for simplicity's sake, I have grandfathered in the optimization which discounts the individuals who seek out spammers unsolicited). If we apply mixed strategies to these outcomes, the rationale behind spamming becomes apparent: there is nothing to be gained from *not* spamming. However, there is also nothing to be lost. Consider that both  $o$  and  $q$  are very low probabilities. Perhaps one in a thousand e-mails yields a response, and one in five thousand results in a counterattack with a non-zero  $d$ . However,  $g$  is an indeterminate monetary sum of enormous potential (generally thousands of dollars over the course of the scam). The value  $d$  is also an indeterminate sum, be it money from having identifying information available to the web (and thus becoming a victim of the very scams the spammer had intended to inflict on others), or the financial cost of the time investment in the construction of the botnet which is lost, or the potential prison sentence if the spammer is vulnerable in that way. An expected utility function helps to elucidate the consideration:

$$E(\text{Solicit}) = (o)(g) + q(-d) + (1-o-q)(0) = og - qd$$

$$E(\text{Not Solicit}) = 0$$

In order for spam to be a fruitful venture, the expectation of solicitation must exceed the expectation of not soliciting. Solving this inequality yields the conclusion  $og > qd$ . In other

Boyles

words, whenever the probability of catching a victim multiplied by the amount which can be garnered from said individual exceeds the probability of engaging a security expert multiplied by the amount of damage that the expert can deal to the spammer, spam will abound.

This suggests several obvious solutions to the troubles created by spammers: first, increase the number of security experts. Giving formal training in the art of digital security is a daunting task, even if the students are ready and willing. Second, decrease the number of people who will fall for such scams. This has been the approach of choice for some time, but making people change bad security habits is at least as hard as compelling them to break any other habit. Third, you could increase the expected damage to a spammer in the event of a counterattack. This option is perhaps the least practicable: shy of deploying teams of assassins, no threat will sufficiently deter a spammer. Finally, simply streamline the way damage is dealt to spammers. In other words, increase the probability that a counter attack will occur.

Large internet fixtures like Google and Wikipedia are aggressively engaged in limiting or eliminating spam from their own data. Many have found that increasing the probability of a counterattack reducing the ability of the botnet to act is the most viable solution. Wikipedia, for instance, moderates its content entirely by *crowd-sourcing*, or consigning the responsibility to protect meaningful data to all the users of Wikipedia. Since the editing process is anonymous, any bot can discretely insert its own text into meaningful content. Luckily, this yields two results. The first is that the bot, being fairly unclever itself, can only insert useless information (like extraneous links), which betrays the nature of its authorship, and can be readily undone by any other reader (Priedhorsky et al. 2007). Second, the IP address of the bot is recorded, and the computer will be banned from further editing Wikipedia entries. If every computer in a botnet is subjected to this process, the botnet is robbed entirely of its usefulness to the spammer, leaving

Boyles

him/her with the unfortunate utility score of  $-d$ . It is thus a less-than-useless expense of time to direct a botnet to attempt to deface Wikipedia.

A solution like Wikipedia's is fine for a major venue like Wikipedia, however the same bots can be directed towards thousands of other wikis built on the same platform as Wikipedia, MediaWiki. Without the security of thousands of willing editors, these wikis are assaulted daily. Thus, most of these and similar services turn to the more corporate approach relied upon by Google for its products. Instead of relying on users to freely write the security rules and exceptions (which is itself an imperfect and dangerous process for reasons completely independent of spammers), it subjects users to a battery of tasks to verify their identity. Unfortunately, this has proven largely ineffectual. Without raising the likelihood of a counterattack or the penalty faced by one, Google has given spammers literally no incentive not to target Google products. Google instead takes less direct approach of developing increasingly complex algorithms to weed out spam while not committing the type II error of blocking meaningful or useful communications, lowering the probability of a solicitation yielding a response. This has the unfortunate side-effect that users are less attuned to the dangers of spam, and thus more receptive when spam does penetrate the filters, increasing both the potential take ( $g$ ) and the probability that the recipient will respond ( $o$ ).

Ultimately, spam is a simple game executed millions of times over every day. And so long as there are enough people receptive to the enticements, and few enough people who can retaliate to them, the business of spam will continue unabated.

## References

Cárdenas, Alvaro A., John Chuang, Jens Grossklags, Chris Hoofnagle, and Svetlana Radosavac.

“An Economic Map of Cybercrime”. (2010). Working Paper.

Carr, Jeffrey, and Sanjay Goel. 2010. “Project Grey Goose Report on Critical Infrastructure:

Attacks, Actors, and Emerging Threats”. GreyLogic White Paper, January 21, 2010.

Priedhorsky, Reid, Jilin Chen, Shyong (Tony) K. Lam, Katherine Panciera, Loren Terveen, and

John Riedl (GroupLens Research, Department of Computer Science and Engineering,

University of Minnesota) (2007-11-04). "Creating, Destroying, and Restoring Value in

Wikipedia". *Association for Computing Machinery GROUP '07 conference proceedings*

(Sanibel Island, Florida).