Fishing for Phishes: Applying Capture-Recapture Methods to Estimate Phishing Populations

Rhiannon Weaver
CERT Network Situational Awareness Group
5000 Forbes Avenue
Pittsburgh, PA 15213
rweaver@cert.org

M. Patrick Collins
CERT Network Situational Awareness Group
5000 Forbes Avenue
Pittsburgh, PA 15213
mcollins@cert.org

ABSTRACT

We estimate of the extent of phishing activity on the Internet via capture-recapture analysis of two major phishing site reports. Capture-recapture analysis is a population estimation technique originally developed for wildlife conservation, but is applicable in any environment wherein multiple independent parties collect reports of an activity.

Generating a meaningful population estimate for phishing activity requires addressing complex relationships between phishers and phishing reports. Phishers clandestinely occupy machines and adding evasive measures into phishing URLs to evade firewalls and other fraud-detection measures. Phishing reports, in the meantime, may be demonstrate a preference towards certain classes of phish.

We address these problems by estimating population in terms of netblocks and by clustering phishing attempts together into *scams*, which are phishes that demonstrate similar behavior on multiple axes. We generate population estimates using data from two different phishing reports over an 80-day period, and show that these reports capture approximately 40% of scams and 80% of CIDR /24 (256 contiguous address) netblocks involved in phishing.

1. INTRODUCTION

This paper estimates the extent of phishing activity on the Internet by collating information from multiple phishing reports. In comparison to other forms of computer crime, phishing depends on explicitly announcing a site in order to attract targets. Online watchdog sites routinely collect this information, providing reports of phishing activity. This information provides us with a collection of events that can be used to estimate the population of phishing activity using an estimation method called *capture-recapture*.

We see this population estimation work as a necessary component for understanding the extent of the threat posed by phishing and other hostile activities to the public health of the Internet. In particular, CERT/NetSA studies phishing as part of a broader examination of *uncleanliness*, which is the likelihood of a particular IP address being used to launch attacks [4]. Since uncleanliness values are calculated by collating reports of hostile activity, we use population estimates to determine how complete our coverage of a particular threat is.

In comparison to other uncleanliness indicators, phishing requires creating and *publicly* advertising a fraudulent site. Phishing reports are therefore, relative to other phenomena, trustworthy reports of an unclean activity. Phishing sites obviously do deceive end users but, unlike DDoSes, which can come from spoofed addresses, or scanning, which can be conducted stealthily, a phishing site is identifiable as such by the owners of the site it impersonates. In addition, since multiple organizations regularly identify and shut down phishing sites, we have access to reports from diverse sources.

The question of estimating population size originally arose in wildlife studies, where the goal was to learn how many animals of a certain species lived in a particular habitat. The method that was devised, capture-recapture estimation, relied on the ability to take two independent random samples of the population. Since that time, capture-recapture analysis has been extended for application to more complex populations and sampling schemes. In general, capture-recapture models can all be considered as variants of a statistical tool called the generalized linear model [12], and the capture-recapture approach is well-suited to population estimation from reports.

However, estimating phishing populations is complicated by the ambiguous relationship between URLs, the machines hosting those URLs and the phishers controlling those machines. In order to provide scalable services, a single domain name (such as `www.ebay.com`) can resolve to hundreds of individual servers located around the world; consequently a phish located on a hosting service may be present on multiple IP addresses. This ambiguity is exacerbated by the phishers themselves as they use increasingly sophisticated phishing tools to evade detection.

In order to compensate for these factors, we cluster records from individual phishing reports into constructs that we term *scams*; a scam is a collection of phishing records which share common attributes, such as address and phishing URL. Once we have grouped phishes into scams, we can examine the populations. To do so, we apply capture-recapture analysis as an exploratory tool for two different population values: *netblocks*, which are aggregations of IP addresses into

APWG eCrime Researcher Summit, October 4-5, 2007, Pittsburgh, PA, USA..

24-bit CIDR blocks and *scams*, which are clusters of related phishing reports.

The rest of this paper is structured as follows: §2 describes the source reports and the methodologies used by these sources to gather data. §3 describes our method for clustering phishing attempts into scams. §4 describes the capture-recapture population estimation technique and the complexities involved in applying capture-recapture to phishing data. §5 applies capture-recapture to IP addresses, and §6 applies capture-recapture to scams. §7 discusses the results of this analysis. §8 discusses previous efforts relevant to this work.

2. DATA

As noted in §1, phishing population estimates are complicated due to the relationship between *phishes*, by which we mean a unified effort by a phisher to deceive *marks* into giving up their identities, and the multiple reports that describe a single phish. In this section, we describe the data we use to analyze phishing populations in order to differentiate phishes from their reported data.

Our data sources are *reports*, which we define as a collection of *records* coming from a single *source*. Each record is a tuple containing the following fields:

- reportdate: the date on which the phish was reported to a source. For this analysis, reportdate is always between January 1 and March 21, 2007.
- target: the institution that the phish imitated.
- url: the URL that marks were instructed to click on in the phishing mail.
- address: the IP address of the site hosting the phish (*i.e.* the IP address of the web server that hosts url).

We use these records as indicators of *scams*. A scam is a construct created from a cluster of multiple similar records; the method for constructing scams is detailed in §3.

We used two reports for our analysis. The report identified as netcraft in this paper is collected by NetCraft, LLC[1]. The report identified as castlecops in this paper was prepared from data collected from the Phishing Incident Reporting and Termination Squad[2] Table 1 summarizes the number of unique values for each field in each report. Due to the data reporting practices of each source, these values are recorded slightly differently in each report.

Records in netcraft are used by a phishing toolbar that validates sites for users. Because the netcraft data is geared toward user protection, it is less concerned with verifying when a site is no longer a phishing site, therefore it does not track how long phishing sites stay active. Each distinct (target, url, address) tuple is a unique record in the report. Figure 1(a) shows the distribution of the number of records observed per /24 CIDR block. As this figure shows, the majority of /24s on the report contain only one record. The top 0.5% (37 netblocks) had 100 records or more, with a maximum of 2,271 records observed within a single 24-bit CIDR block.

[1] `http://www.netcraft.com`

[2] as described by PIRT, "PIRT is a global phishing termination operation launched by CastleCops and Sunbelt Software. PIRT is operated at `www.castlecops.com`, a volunteer security community focused on making the Internet a safer place."

Records in castlecops come from a volunteer-based reporting site; visitors to the site may use the form to enter information about any phishing sites they have encountered. CastleCops also maintains a message board where volunteers post and track phishing activity. In most cases, members keep records open and revisit links *until* the links are removed, which yields approximate birth and death dates for records. However CastleCops distinguishes between records by address value, as a consequence, multiple different phishes hosted on the same site may be reported as a single incident.

Figure 1(b) shows the distribution of the number of records observed per /24 netblock for castlecops. The top 6.9% (45 netblocks) had 100 records or more, with a maximum of 1,786 records observed.

Initial examination of castlecops yielded a total of 98 unique targets. However, target identifiers were occasionally ambiguous due to the presence of the *Rock Phish* toolkit[3]. Rock Phish uses an unusual dynamic that imitates multiple targets on a single phishing site. Hand examining the Rock Phish records and cross-referencing the results with targets in netcraft revealed at least 14 added targets within the Rock Phish records, for a total of at least 112 unique targets.

As the proportion of addresses to /24 CIDR blocks indicates, the majority of these CIDR blocks contain, at most, a single IP address hosting a phishing site. Consequently, we can reduce the IP addresses to their /24 CIDR blocks in order to reduce clustering complexity without significantly sacrificing precision. To that end, we will refer to addresses as *netblocks* to indicate the 24-bit CIDR block prefix containing the address in question.

3. SCAM CLUSTERING

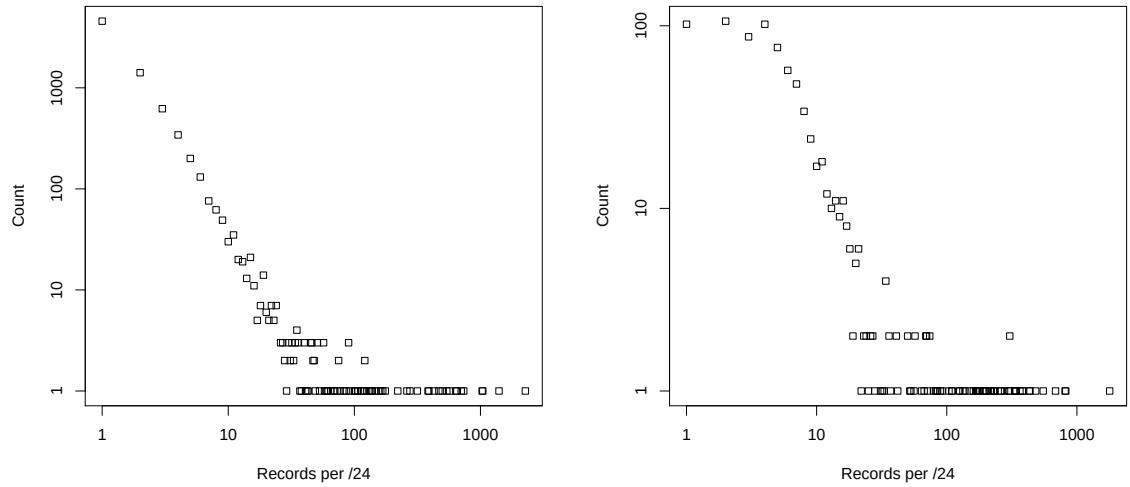
Recall in §2 we defined a scam as a construct that approximates a single phish by clustering together similar records. In this phase of the work we describe the methods used to create scams.

We cluster records into scams in order to compensate for multiple records received for the same phish. Even if both reports receive the same observable data, a single phish can be represented as a different number of records in each report. Often, similar reports are made very closely in time, using variations of the same URL structure. For example, consider the following URLs for eBay targets on a single IP address as reported to netcraft:

```
http://signin.ebay.com.h9lwhws.03iana.com/.../eBayISAPI.php
http://signin.ebay.com.n73vljf.03iana.com/.../eBayISAPI.php
http://signin.ebay.com.drkzzgo.03iana.com/.../eBayISAPI.php
http://signin.ebay.com.rggtjlz.03iana.com/.../eBayISAPI.php
http://signin.ebay.com.49smxz6.03iana.com/.../eBayISAPI.php
http://signin.ebay.com.0ysgfpc.03iana.com/.../eBayISAPI.php
```

The report times of each of these records were all within 30 seconds of each other. Coupled with the similarly structured URLs and the equal IP address, we can reasonably assume that these records were not generated independently, and that they may all be representative of the same underlying phishing action. To compensate for these dependencies, we introduce a naive clustering method for grouping phishing attempts by target, url and address to detect scams. Ideally

[3] Robert McMillan, "Rock Phish blamed for surge in phishing", IDG News Service, `http://www.infoworld.com/article/06/12/12/HNrockphish_1.html`, retrieved 2007/05/19.



(a) netcraft (b) castlecops

Figure 1: Distribution of the number of phishing records seen per /24 for each report (Jane 1 – Mar 21, 2007)

Report	Records	Count of target	scam	address	/24s	/16s	/8s
netcraft	32,680	435	16,566	7,539	5,317	2,763	106
castlecops	20,816	112	2,533	828	646	561	69

Table 1: Summary of report attributes

a scam will approximate the actual distribution of unique phishes conducted at any time.

This section is structured as follows: §3.1 describes the metrics used to generate the scam. §3.2 describes a simple clustering method.

3.1 Covariate Distance Metrics

We develop the scam construct by examining data within the reports. We wish to develop a method that will cluster reports into groups that are similar. If clusters are easily identifiable and separable, each can be labeled as a distinct scam, and the population estimate can then be expressed in terms of number of scams. In order for scam clustering to be useful, it must demonstrate two properties:

- *specificity*: the ability to recognize separate scams as signatures from separate individuals;
- *low-cost action*: a high correlation between reports in a scam and the ability to shut them down.

In practice we will not be able to measure these properties directly, as we do not have ground truth for the relationship between scams and phishes. However, we can associate these properties with a set of covariates based on the fields in the report records. Recall the variables introduced in §2; the covariates of url, target, and address can each be evaluated in terms of specificity and low-cost action: phishes that imitate the same targets with similar URLs are likely to be related by controller or phishing kit. Similarly, phishes that have closely related IP addresses may be related not only by deployment of a large-scale phishing kit, but by the hosts and policies of the host's netblock, and will be easier to block *en masse* via blacklisting IP ranges or alerting ISPs or system administrators.

3.2 A Simple Clustering Method

Records from both reports were aggregated into a set of scams using a simple clustering method based on target, url and address. We define the *IP distance* $d_{\mathsf{IP}}(a, b)$ for two records with address as in the same netblock as the absolute value of the difference of the final 8 bits of the address value. We define the *URL distance* $d_{\mathsf{URL}}(a, b)$ for two records as the Levenshtein distance [10] between the url values of a and b. To cluster records, we then use the following algorithm:

1. Let N_{ij} be the number of records in netblock i where target $= j$.
2. Start with the record k_0 that has the earliest reportdate, and create a scam $S_{ij}^0 = \{k_0\}$.
3. For $n = 1, ...N_{ij} - 1$:

(a) Let $S_{ij}^0, ..., S_{ij}^M$ be the collection of scams for netblock i and target $= j$ before record k_n is added.

(b) Set $m = 0$. While $m \leq M$ and k_n is not matched to a scam:

i. if for all $k_r \in S_{ij}^m$ we have,
$d_{\mathsf{IP}}(k_n, k_r) < 5$ and
$d_{\mathsf{URL}}(k_n, k_r) < 20$,
add k_n to scam S_{ij}^m.

(c) if k_n is assigned a scam when all scams have been examined, create $S_{ij}^{M+1} = \{k_n\}$.

Note that the distance thresholds were arbitrarily chosen in order to quickly group related records. Cross-report matching was done by comparing the url values within scams. If the records for scam S_a in netcraft contained a URL exactly matching url in the record of scam S_b in the castlecops, then the two scams were considered the same scam. After obtaining an initial list of cross-report matches, scams within the *same* report were merged together if they were shown to match a common scam in the other report. For instance, if scams S_{a1} and S_{a2} in netcraft were both shown to match a single scam S_{b1} in castlecops, then the scams S_{a1} and S_{a2} were merged together to form a new scam $S_{a1.2}$.

Table 1 shows the number of unique scams generated for each report; as this table shows, the 32,680 records in netcraft cluster into 16,566 scams, and the initial 20,816 records for castlecops cluster into 2,533 scams. The resulting scam and netblock populations are used throughout the remainder of this paper for population estimation and analysis.

4. CAPTURE-RECAPTURE POPULATION ESTIMATION

We now focus on the methodology used to estimate the phishing population, *capture-recapture estimation*. In this section, we describe the basics of capture-recapture population estimation and the factors which may impact the estimate. This section is structured as follows: §4.1 reviews the basic capture-recapture approach. §4.2 addresses the specific issues involved in applying capture-recapture analysis to report data for complex populations.

4.1 Basic Capture-Recapture

The purpose of capture-recapture estimation is to count the number of individuals when a census cannot be taken due to cost or logistical difficulty. Because phishers are necessarily interested in obfuscating their population, capture-recapture is well suited for addressing this problem.

Capture-recapture works by using repeated sampling of the population. Although a true count of the population cannot be obtained in this fashion, an estimate can. In a simple capture-recapture experiment, an initial sample S_1, with an observed size of N_1 individuals, is captured, marked, and released back into the habitat. At a later date, after the original sample has had time to re-integrate into the population, experimenters obtain a second sample S_2 of N_2 individuals. The *intersection* between S_1 and S_2 is the set of individuals common to both samples. We calculate the *overlap* of S_1 and S_2 as the number of individuals in the intersection, recorded as $M_{1,2}$. The value $\hat{p} = \frac{M_{1,2}}{S_2}$ is then the observed proportion of marked individuals in S_2. Assuming that both samples S_1 and S_2 are simple random samples from a homogeneous population, $\hat{p}$ is an estimate of the total proportion of marked individuals in the habitat. Thus, estimate the total population size, $\hat{N}$ as

$$\hat{N} = \frac{N_1}{\hat{p}} = \frac{N_1 N_2}{M_{1,2}}. \quad (1)$$

The capture-recapture method can be generalized to $K > 2$ independent samples by using log-linear models [11, 8].

Capture-recapture can also be employed when the population is comprised of individuals recorded in reports from multiple sources. In this case, each report is treated as a sample in a multiple-recapture experiment. An individual appearing on a report has been "marked" by that report's source, and overlaps between sources are used to estimate the total population. The method has been employed in a variety of settings where report data arises, for example epidemiological studies of disease prevalence [19], census data [7], and error-checking in software development [2].

4.2 Adaptations for Complex Populations

The simple model for $\hat{N}$ from Equation 1 depends upon strong assumptions about both the population and the sampling method. To ensure N is the same for both samples, the population must be *closed*, meaning no births or deaths occur in the interim period. In addition, the the population is assumed to be *homogeneous*, to the extent that differences among individuals of interest do not make them more or less likely to be sampled. Finally, the samples themselves are assumed to be *independent*, simple random samples; that is, individuals all have the same probability of capture, and initial capture does not affect the probability of recapture.

These assumptions are questionable when applying capture-recapture to phishing, because phishing scams comprise a diverse, dynamic population that is measured heterogeneously among sources, and that reacts to observation. Individual URLs are constantly entering and leaving the population due to phishers building new sites and to phish fighters taking these sites down. We may presume that scams themselves are heterogeneous; for example, scams targeting large, powerful entities like eBay, PayPal or the Bank of America may be aggressively targeted and reported due to the active interests that the targets have in preventing phishing. Finally, differences between methods for compiling reports may also give the reports biased views into the population. Therefore, a capture-recapture framework must be robust enough to handle complexities such as:

- Heterogeneity: differences among phishers or capture methods that affect the probability of capture. This may include the sophistication and subtlety of a phishing site, or the different reporting and classification methods used by distinct sources.
- Locality: the tendency for individuals to cluster within the population, and of sources to sample only fractions of the whole. In phishing, an example of such a tendency would be for multiple phishers to use the same host address.
- Open populations: *births* and *deaths*, in this context meaning a change in the true population occurring during the sampling periods. In phishing, such phenomena would include the the introduction or removal of phishing sites.

		In netcraft		
		Yes	No	
In castlecops	Yes	493	153	646
	No	4,824	?	?
		5,317	?	?

Table 2: Summary of netblocks reported to house phishing URLs by netcraft and castlecops from January 1 through March 21, 2007.

- Trap effects: effects from the capture that affect recapturing a particular target. A potential trap effect has been observed in botnets by Ramachandran *et al.* [16], where bots known to be on blacklists become less attractive to bot owners. Phishers may have a similar motivation, and consequently may demonstrate less interest in a host once it has been identified as a phishing site.

Heterogeneity and locality can be accounted for by measuring *covariates* for individuals or source reports. A covariate is a measurement taken on an individual that is hypothesized to affect that individual's probability of capture (*e.g.*, size, weight, capture location, speed). Covariates stratify the population, allowing for differences in the probability of capture among strata.

Heterogeneity and locality are especially troubling with third party sources, because reporting methods often cannot be accounted for in detail. When unknown report effects are present in conjunction with individual differences, it is impossible to discern between stratified population estimates that truly reflect different population sizes among strata, and estimates that reflect how the records were compiled, when no further information is available.

To account for open populations, capture-recapture models are adapted to study birth and death rates over time. A simple method is to break the collection time into a set of stable periods and then allow "births" and "deaths" at the beginning and end of each period. Capture-recapture estimates are obtained on the number of births at the start of each period, and deaths are calculated based on observed death times or models for typical lifespans.

When trap effects can be measured or studied *a priori*, the population estimates can be adjusted to account for these effects, although this leads to increased variability in the estimates due to the layer of uncertainty associated with the trap effects themselves. In the following sections, we analyze the records from netcraft and castlecops using exploratory capture-recapture methods with these complications in mind.

5. PHISHING BY NETBLOCK

Table 2 shows the overlap between netblocks in netcraft and castlecops. We now apply capture-recapture estimation to netblocks in these reports in order to estimate the population not recorded. In this section, we examine the population of netblocks that are hosting phishing activity, breaking the results into different strata by using *scam activity* over time as a covariate.

The goal of this analysis is not only to estimate the population of netblocks containing phishing sites, but also to characterize the distribution of phishing scams within netblocks. In order to do so, we characterize netblocks scams both by the *frequency* of scams, but also their *duration*.

We define $\mathsf{scams}(j)$ as the set of distinct scams hosted in netblock j over the observation period, using $|\mathsf{scams}(j)|$ to denote the cardinality of the set. We define $T(\mathsf{scams}(j))$ as the set of scam durations (measured in days) for $\mathsf{scams}(j)$, and then let $\bar{T}(\mathsf{scams}(j))$ be the average length of observed scams with $\mathsf{scams}(j)$. We then divide the blocks in our dataset into four strata, based on two categories for $|\mathsf{scams}(j)|$ and $\bar{T}(\mathsf{scams}(j))$.

- Isolated: these blocks have a limited number of scams, and what number are observed are also short lived ($|\mathsf{scams}(j)| < 10$ and $\bar{T}(\mathsf{scams}(j)) \leq 3$).
- Persistent: these blocks have a limited number of scams, but those scams have a long duration ($|\mathsf{scams}(j)| < 10$ and $\bar{T}(\mathsf{scams}(j)) > 3$).
- Bursty: these blocks have a large number of relatively short-lived scams ($|\mathsf{scams}(j)| \geq 10$ and $\bar{T}(\mathsf{scams}(j)) \leq 3$).
- Corrupt: these blocks have a large number of long-lived scams ($|\mathsf{scams}(j)| \geq 10$ and $\bar{T}(\mathsf{scams}(j)) > 3$).

The intuition for this classification is that we can characterize the duration of a scam by the uncleanliness [4] of a netblock. Netblocks which are controlled by phishers will likely have the same scam reported repeatedly over time, but may be less vulnerable to new scams, which would characterize behavior in the Corrupt stratum. Poorly administered netblocks may show both repeated reports of the same scam, and regular infection with new scams, characteristic of the Persistent stratum. Vulnerable netblocks are those that repeatedly show short-term, new infections with time, due possibly to an unpatched exploit, characteristic of the Bursty stratum. Clean networks may show short bursts of scams coinciding with vulnerability detections, but will be very quick to patch the exploit and purge the network of the affected sites, characteristic of the Isolated stratum.

Classifying netblocks also allows for the possibility of different capture profiles among these strata. Recall from §4.2 that trap effects can impact the population estimate in capture-recapture. We hypothesize that persistently compromised netblocks may consist of a small percentage of the overall population, but due to their activity level, they will be more visible to the source reports. Conversely, netblocks that are compromised for short periods may be more difficult for either report to find.

Figure 2 shows the scam activity in representative samples of each stratum over the observation period. In this figure, each distinct netblock is represented by a line that is present when at least one scam is reported and absent when no scams are reported. The line height is proportional to the number of distinct scams reported by that netblock during the observation period.

Note the gaps that appear in the Persistent and Isolated strata after February 6. These gaps could represent a change in administrative policy for a large number of netblocks, or a change in report methodology for one or both of the reports. To compensate for this difference, we divide the Isolated and Persistent strata into two sub-strata demarcated

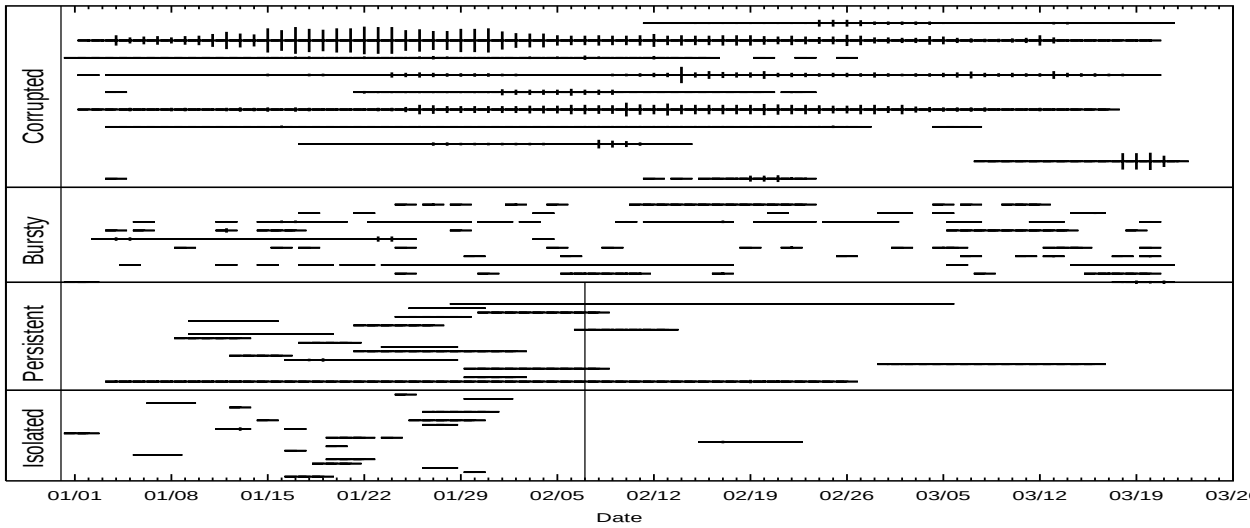


Figure 2: Scams per representative netblocks for observation period divided by strata. Note the pronounced drop in activity in the lower strata after February 6.

	Characteristics			Capture Profile			Population Estimate		
	Date Range	$\|\mathsf{scams}(j)\|$	$\bar{T}(\mathsf{scams}(j))$	netcraft	castlecops	Both	$\hat{N}$	Rate	95% CI
Isolated-1	01/01-02/06	< 10	≤ 3	1,410	25	24	1,469	0.96	(0.71,0.99)
Isolated-2	02/07-03/21	< 10	≤ 3	279	48	20	670	0.45	(0.27,0.66)
Persistent-1	01/01-02/06	< 10	> 3	2,669	227	207	2,927	0.92	(0.85,0.95)
Persistent-2	01/01-03/21	< 10	> 3	726	210	111	1,374	0.60	(0.50,0.70)
Bursty	01/01-03/21	≥ 10	≤ 3	117	43	43	117	1.00	(0.89,1.00)
Corrupt	01/01-03/21	≥ 10	> 3	116	93	88	123	0.98	(0.94,1.00)
Total				5,317	646	493	6,680	0.82	(0.62,0.91)

Table 3: Capture-recapture estimates for six strata of netblocks.

by February 6, and examine capture-recapture results within each substratum separately.

Table 3 shows the results of applying capture-recapture to the six strata of netblocks described by the four scam characteristics and by splitting the Isolated and Persistent strata over time. The capture profile gives the total number of netblocks marked by each report marginally, and the number of netblocks marked by both. The population estimate gives a point estimate of the population N for each stratum, as well as the *capture rate* and a 95% confidence interval for the capture rate, based on the method outlined in Lohr [11]. The confidence interval for the total was obtained using the Bonferroni correction for the six individual strata [17].

Overall, it appears that these two reports have identified approximately 80% of netblocks with active phishing sites during the observation period. As hypothesized, the more visible netblocks with consistent activity over time had higher capture rates. Also, the shift in behavior from January to February and March appears to be a result of fewer records in netcraft, rather than a slowdown in phishing activity. The capture rates for Isolated and Persistent strata drop dramatically between the two time periods.

Tracking the population of netblocks with phishing activity is an easier task than monitoring and characterizing the scams themselves. In the next sections, we explore methods for clustering related records into scams, and estimating scam birth and death rates in the presence of heterogeneity in report methods and scam characteristics.

6. PHISHING BY SCAM

The results of matching scams across reports yielded a total of 18,593 scams, 506 of which were recorded on both reports. Table 4 shows the breakdown of counts. Although the counts from Table 4 could be used in a simple capture-recapture population estimate using Equation 1, the estimate has little value because the scam population does not adhere to the assumptions of the model that Equation 1 defines. Scams are heterogeneous and dynamic, and we must address these complications in a capture-recapture analysis.

In this section, we use capture-recapture estimation as an exploratory method to study the population of scams. This work is a preliminary diagnostic step in constructing a statistical model that will incorporate more robust data, prior knowledge and expert analysis to clarify the effects of list bias and heterogeneity on population estimates, as well as the effect of open populations. In §6.1, we use simple graphs to assess the impact of heterogeneity, locality, open popula-

		In netcraft		
		Yes	No	
In castlecops	Yes	506	2,027	2,533
	No	16,060	?	?
		16,566	?	?

Table 4: Summary of scams recorded on netcraft and castlecops from January 1 through March 21, 2007.

tions and trap effects in the scam population. In §6.2, we estimate births by day for scams using the simple capture-recapture method for each day, and in §6.3, we again use the simple capture-recapture method to estimate the proportions of the population characterized by eight different types of scams over the observed time frame.

6.1 Assessing Complications

Because we now turn our attention to the scams themselves as opposed to netblocks with scam activity, we must focus more carefully on birth and death rates for scams, as well as describing covariates for scams that can be used to stratify them into meaningful subgroups. The population of scams is more diverse than that of /24s, the report reports may or may not be biased, and the scams themselves are transient. In this section, we attempt to examine and diagnose the effects of heterogeneity, locality, open populations and trap effects for scams.

6.1.1 Heterogeneity

For each scam, we record information on the scam size and type. Size refers to the number of records that comprise the scam, and the type is a reduced categorization of the target, classified into one of 10 strata named by industry. Figure 3 shows a plot of the density of the number of records per scam for each of the three capture profiles. Table 5 lists the breakdown of scams by type, and the percentage of each capture profile that consists of the listed type. The majority of the observed scams (85%) fall into the top three strata of bank, ebay and paypal.

We use these covariates to characterize strata of interest for a breakdown of capture rates in §6.3, and to study factors influencing lifetime estimates in §6.1.3. We hypothesize that scams consisting of many records are hallmarks of more organized phishing groups such as the Rock Phish group, or of black market distribution kits, and it would be advantageous to have the ability to cluster these records into scams using only the information from reports, as opposed to more detailed information based on surveillance and expert tracking.

6.1.2 Locality

The clustering method described in §3 is a way of accounting for locality on the individual level by re-defining an individual as a set of related records. But this does not account for the tendency of reports to sample parts of the whole. Locality in this sense can be seen as a version of report heterogeneity; the places where reports are likely to find individuals is a function of their reporting methods. We assume that each report obtains records independently from the other, and that the reports cover a broad range of IP space. We note that both reports are biased toward

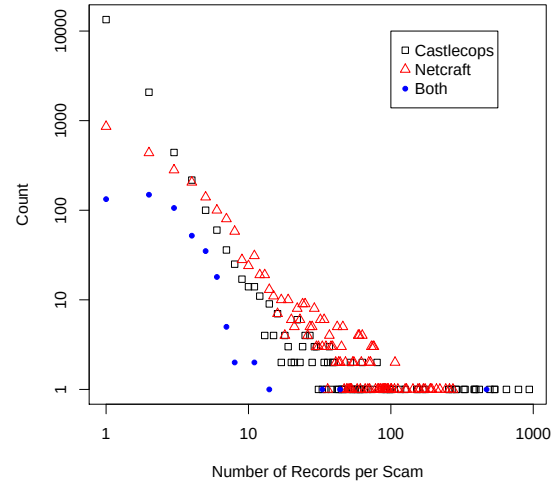


Figure 3: Distribution of the number of records per scam for three capture profiles.

		Percent of Scams		
Type	Count	netcraft	castlecops	Both
bank	8,412	42.0	63.0	37.2
ebay	4,187	24.7	6.4	12.5
paypal	3,390	18.5	18.4	29.2
creditcard	725	4.1	2.6	4.3
retail/business	425	2.0	4.9	8.1
creditunion	407	2.2	1.8	2.8
insurance	345	2.1	0.3	0.9
internetservice	219	1.3	0.5	1.2
various	494	2.3	2.0	3.8
unknown	75	0.4	0.1	0.0
Total	18,593	100.0%	100.0%	100.0%

Table 5: Summary of scam types recorded on netcraft and castlecops from January 1 through March 21, 2007.

the Western, English-speaking world as opposed to Asian languages. This pertains to the kinds of targets and victims the phishers are scamming, not to where the URLs are housed. We are currently working with Asia-Pacific sources to catalog a list of Asian language phishes as well.

One way to check for report agreement is to study the distribution of population covariates across capture profiles. Disparate distributions give an indication of the tendency of one report to capture certain kinds of records more easily in comparison with another report. Similar distributions indicate that both reports capture the same kinds of records, which could be a sign either that each report is sampling independently over the population, or that each report is biased in the same direction. The percentages in Table 5 are similar for the three different capture profiles, but do suggest some report heterogeneity: castlecops appears more centered on banking scams, with comparatively fewer reports in the ebay stratum. Modeling this heterogeneity is beyond the

scope of the preliminary analysis, but it is a feature that we will study closely in future methodological developments.

6.1.3 Open Populations

To account for the short term stability of phishing scams, we separate the time frame into days, and study births and deaths of scams for each day. This requires observations of life spans for each scam. Since castlecops records duration, this duration was used as the life span for scams appearing in castlecops. An approximate duration was compiled for scams in the netcraft consisting of more than one record by using the minimum and maximum report times of the records in the scam.

For scams appearing in both reports, if netcraft's scam had duration 0, the life span recorded by castlecops was used. If netcraft's scam had multiple records whose span was larger than the duration listed by the castlecops the larger duration was used.

For netcraft records with duration 0, a lifetime was imputed using the distribution observed for lifetimes of castlecops scams. The breakdown of lifetime by type (Figure 5) indicates that the distribution of lifetimes appears relatively consistent. Distributions are very similar for top three target types (bank, ebay, and paypal), with median values fluctuating only slightly across types.

Figure 4 shows the distribution of lifetimes for the castlecops scams. Comparing with the data appearing in both reports, the distributions are very similar, suggesting that extrapolating the castlecops distribution to 0-duration netcraft records is a reasonable course of action. We note that this distribution is truncated toward shorter values due to the sampling window: scams appearing later on in the time frame may have "Last known active" dates of March 21, as opposed to observed deaths.

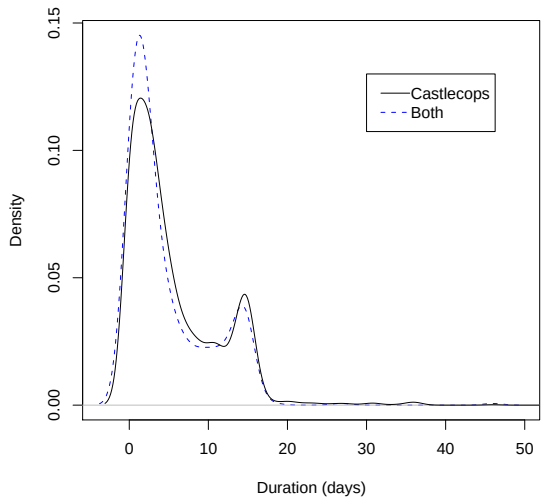


Figure 4: Distribution of the active lifetime for castlecops records, and for castlecops records that also appeared on netcraft.

Based on this analysis, for each netcraft record with duration 0, a lifetime in days was imputed with a random draw

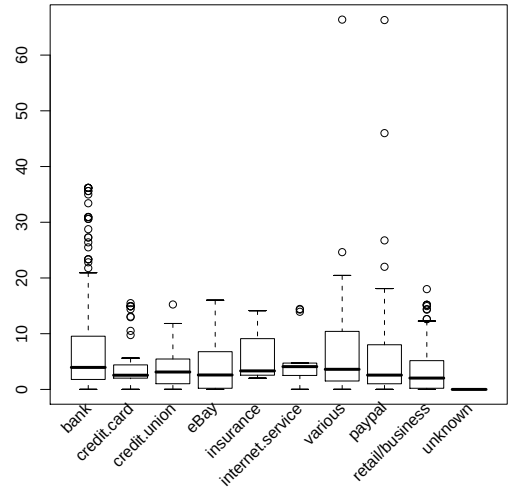


Figure 5: Distributions of lifetime duration for 10 different scam targets.

from the discrete distribution using the observed frequencies from castlecops.

6.1.4 Trap Effects

In the case of reports of phishing using independent sources, we are concerned with trap deaths. If one report uses another as a source, it is reasonable to assume that a record marked in the first report has a higher probability of showing up on the second. But if each report acts independently, we might expect that a record appearing on the first report has a lower chance of appearing on the second because the site will be more likely to be taken down or moved. Figure 6 shows a histogram of the time lag (castlecops - netcraft) of the first instance of record for the 506 scams appearing in both reports. The distribution has a mode at 0, with 50% of lag times falling within 12 hours, spread relatively evenly between netcraft and castlecops in which report logged the scam first. In only 36 cases (approximately 7%), castlecops lagged behind netcraft for 5 days or more. netcraft lagged behind castlecops for more than 3 days only once.

Comparing the lag time distribution with the average lifetime, it appears that lags between the two reports are generally shorter than the average or median lifetime of scams. In this sense, we can assume that the effect of trap deaths on the population is small.

6.2 Scams over time

Figure 7 shows the results of capture-recapture analysis on births by day for the scam data. The dotted line is the result of applying capture-recapture independently for each day, using the new scams reported on that day as the observed population. The day-by-day estimates have large variability due to small numbers of reports per day, and to small overlaps relative to marginal list sizes per day. Because overlaps appear in the denominator of the capture-recapture calculation (Equation 1), estimates can be unstable when the overlap is small relative to the total number of records

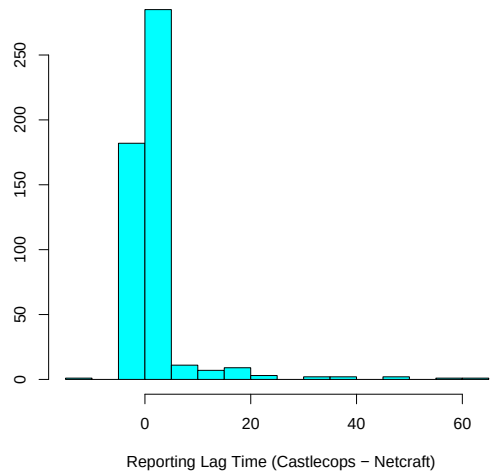


Figure 6: Distribution of the lag time (castlecops - netcraft) of the first instance of record for scams appearing in both reports.

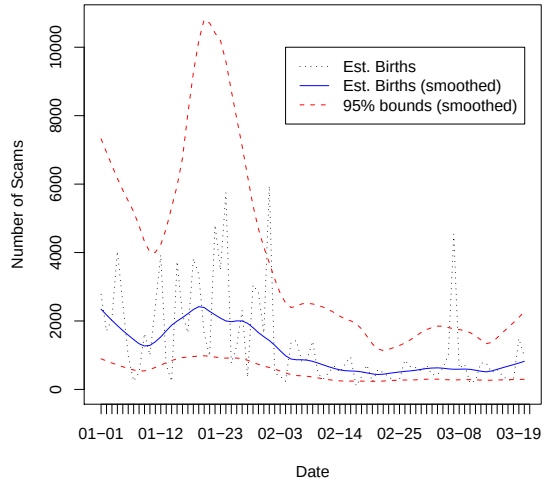


Figure 7: Estimate of the total number of scam births per day.

on each individual list.

Part of the large variability in scam population estimates is also due to the naive method for clustering. For example, the large spike on March 7 is due to the detection by castlecops of a large number of Rock Phish reports spread across several netblocks with URLs and targets that, although similar, fell outside the thresholds for the clustering metric. The fact that these reports were treated as independent inflated the number of observations for castlecops on that day that were not logged by netcraft also inflating the population estimate. A similar spike occurs on January 4 due to a large number of unclustered eBay records appearing on the same netblock in netcraft. The url values for these reports had been injected with long strings of digits that overwhelmed the simple Levenshtein distance metric for clustering.

In light of the naive clustering method and the large variability inherent in treating each day independently, we use local weighted polynomial smoothing [3] to smooth the observed estimates. This yields the solid curve in Figure 7, which displays a general trend for births by day that takes into account the trend on nearby days. The confidence bands (dashed lines) are obtained by applying the same polynomial smoother to the 95% confidence intervals generated by the individual, unsmoothed days. Using the smoothed estimate, we obtain a total of approximately $88,612$ scam births over the 80-day observation period, for an overall capture rate of approximately 21%, given that $18,593$ of those scams were reported. The capture rate is worse when activity is high in January (17%) and stabilizes to approximately 31% during February and March.

Using the smoothed estimate for the number of births per day, we impute lifetimes for the unobserved scams using the distribution from §6.1.3 [4], and use these to obtain estimates of the total number of active scams per day. Figure 8 shows the result.

As with the analysis by netblock, we see a marked difference in the behavior of the two curves from the month of January through February 6, and then from February 7 through March 21. As the birth rate stabilizes, the estimated number of active scams per day decreases, but this could also be a result of shifting new scams to a new set of netblocks that is not tracked well by either report. Counting active scams by day also gives more weight to longer scams; a scam that is logged by either of the reports over a period of k days is observed k times.

6.3 Scams by type

Now we examine capture rates and population estimates for various strata of scams. In this section, we implicitly use birth estimates from §6.2, and report populations for scam strata as percentages. We are interested in phishing kits for eBay and paypal, Rock Phish scams, isolated scams that have unusual targets, and non-clustered scams that may be part of something larger. For this analysis, we will restrict Rock Phish scams to bank targets; in castlecops, the majority of records specifically labeled as belonging to Rock Phish were banks. Let $R(j)$ be the number of records observed for scam j. We use $R(j) \geq 5$ for ebay and paypal as an indicator for phishing kits, and $R(j) \geq 5$ with bank targets as an indicator of Rock Phish scams, which we classify as Kits.

When $R(j) < 5$, we label the three prevalent types (bank, ebay and paypal) as Parts, suggesting that these types tend to appear in larger clusters than have been observed. We label unusual targets as other, consisting of all target types from Table 5 that are not eBay, Paypal or banks (approximately 15% of all scams). Isolated scams are characterized

[4]This distribution is optimistic for scams that are not logged by either report, but may be valid if these reports are linked with others that are observed, or they are observed by other reports

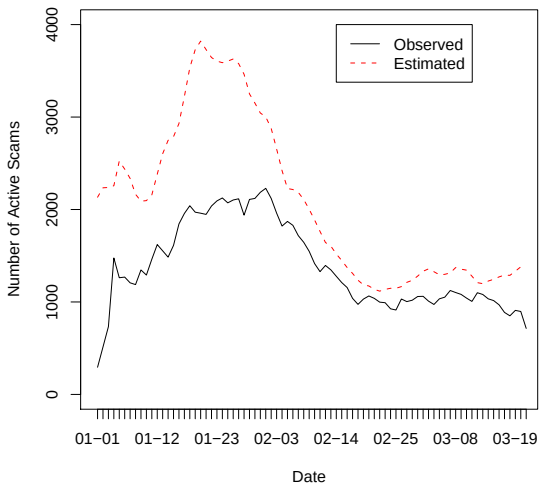


Figure 8: Number of active scams per day, observed (solid line) and estimated (dotted line).

by $R(j) < 5$ with an unusual target. Lastly, to account for the disparate activity with time, we again stratify the sample into two date ranges, as in §5. Based on the diagnostics from §6.2 and examination of the data, we also agglomerated 590 unclustered eBay scams from January 4 into a single scam, before performing the capture-recapture analysis on these strata.

The chosen covariates yield 16 strata over which to perform capture-recapture estimation. Results are shown in Table 6. For each stratum, we list both the total number of observed records and the total number of scams. We use capture-recapture to obtain an estimate of the percent of the scam population that belongs to each stratum (%Scams). Multiplying by the average number of records per scam and re-normalizing yields an approximation of the percent of records appearing from each scam ($\approx$ %Recs). We also give approximate rates of capture for each stratum with associated 95% confidence intervals. These proportions and capture rates are based on the assumptions of report independence and lack of report bias, and we discuss the results with these assumptions in mind.

A striking feature in Table 6 is the estimates for suspected Rock Phish scams. Despite comprising approximately 45% of all records observed on the two reports, the percentage of scams is much smaller. These scams are ostensibly reaching a larger audience due to the number of URLs and the amount of spam emails required to saturate the reports, but they still are characterized by large clusters of very similar URLs closely related in IP space. This suggests that local blocking of IP addresses or domains may be effective in shutting down many sites. But the relationship is not only a spatial one. With time, we expect to see these clusters moving via fast flux[5] mechanisms that traverse larger distances of address, target, and url. Because of the naive clustering methods, we expect that many of these scams are reported as Parts for the bank, ebay and paypal strata. Also, as preference for social networking sites grows, we expect to see an increase in other kits or fast flux isolated phish.

Despite the high degree of clustering in Rock Phish scams, we also note that they have the lowest capture rates out of the four types of kits. This could be a result of the tendency of castlecops to seek out or more aggressively report Rock Phish, but the low overlap even after clustering suggests the presence of many unseen clusters. Ebay and PayPal scams appear to be tracked well by these two reports, with approximately one in three isolated scams being reported, and clusters of kit activity reported more accurately. Rates of capture appear relatively stable across the two time periods, with slight increases as the birth rate decreases.

7. DISCUSSION

When we compare the population estimates and capture rates for netblocks against scams, we immediately note that we have a far higher capture rate for netblocks than we do for scams. According to Table 3, the reports cover approximately 80% of extant phishing netblocks, while we cover approximately 40% of scams. This address clustering would appear to indicate that the majority of phishes are concentrated within a limited number of bad actor netblocks, which are hosting multiple scams.

Any observations on scam population are dependent on the accuracy of the clustering algorithm. The current implementation is a naive implementation intended for exploratory analysis. However, since URLs and addresses have structure, a more sophisticated clustering algorithm may more accurately group scams. This is particularly a concern with scams that use dynamically generated URLs.

Finally, our results are generated using reports from two sources and which may have distinct relationships and biases themselves. With the current data we are not certain, for example, whether the high number of reports for PayPal is a result of a large number of phishes for PayPal or aggressive hunting for PayPal phishes.

8. PREVIOUS WORK

Thomas and Martin [18] describe the fundamental mechanics of the underground economy (including phishers); of particular relevance to this work is the emphasis that the underground economy does not rely on privacy technologies and largely relies on public and observable services. Moore and Clayton [14] study the effective lifetime of phishers by examining data from the PhishTank[6] dataset.

Outside of phishing, Moore *et al.* [13] estimate spoofed denial of service activity by examining *backscatter* (responses to spoofed IP addresses) in unpopulated network blocks, while Ramachandran and Feamster [15] examine spammer lifetimes. Ramachandran and Feamster's work is similar to ours in that it finds that majority of spammers appear within a limited region of IP addresses, however their work is based around direct-observation. Several techniques have been developed for observing botnet populations by tracking IRC sites, DNS traffic and other indicia [5, 6, 9]. Abu Rajab *et al.* [1] discuss many of the mechanical concerns associated with estimating botnet populations. In comparison to these

[5]Spamhaus FAQ, "What is fast flux hosting", `http://www.spamhaus.org/faq/answers.lasso?section=ISPSpamIssues`, retrieved 2007/05/21

[6]`http://www.phishtank.com`

	Characteristics			Capture Profile					Population Estimates			
	Date	$R(j)$	Type	Recs	Scams	netcraft	castlecops	Both	%Scams	≈%Recs	Rate	95%CI
Parts		< 5	bank	5,649	4,255	3,766	565	76	35.66	16.56	0.15	(0.10,0.22)
		< 5	ebay	3,365	2,629	2,592	58	21	9.12	4.08	0.37	(0.18,0.60)
		< 5	paypal	2,481	1,986	1,892	135	41	7.94	3.47	0.32	(0.19,0.48)
Isolated	01/01-	< 5	other	1,815	1,427	1,307	148	28	8.80	3.91	0.21	(0.11,0.35)
Kits	02/06	≥ 5	bank	18,447	552	165	410	23	3.75	43.83	0.19	(0.09,0.34)
		≥ 5	ebay	2,417	71	69	10	8	0.11	1.31	0.82	(0.30,1.00)
		≥ 5	paypal	1,765	76	59	39	22	0.13	1.06	0.72	(0.42,0.95)
		≥ 5	other	346	47	26	32	11	0.10	0.26	0.62	(0.26,0.94)
Parts		< 5	bank	4,195	3,250	2,949	369	68	20.38	9.20	0.20	(0.14,0.29)
		< 5	ebay	1,121	879	827	82	30	2.88	1.28	0.39	(0.22,0.59)
		< 5	paypal	1,661	1,251	1,083	225	57	5.45	2.53	0.29	(0.19,0.42)
Isolated	02/07-	< 5	other	1,521	1,181	1,055	185	59	4.21	1.90	0.36	(0.24,0.50)
Kits	03/21	≥ 5	bank	6,230	269	124	166	21	1.25	10.13	0.27	(0.14,0.48)
		≥ 5	ebay	134	19	11	12	4	0.04	0.10	0.56	(0.13,1.00)
		≥ 5	paypal	506	77	36	69	28	0.11	0.25	0.87	(0.59,1.00)
		≥ 5	other	210	35	16	28	9	0.06	0.13	0.70	(0.27,1.00)

Table 6: Capture-recapture estimates for sixteen strata of scam types.

analyses, our work is focused on collating results from multiple sources and developing models that can compensate for those biases.

9. CONCLUSIONS AND FUTURE WORK

In this paper, we have applied a population estimation technique, capture-recapture, to reports of phishing population in order estimate the extent of phishing present on the Internet. Our work shows that the two reports used cover about 80% of the netblocks and 40% of the extant scams. Our results are partly a function of the clustering algorithm applied to the data, however the initial analysis implies that phishing sites are clustered in specific netblocks demonstrating bad behavior. The results are a preliminary exploratory analysis to gain insights into the important features of the data and to guide development of a more rigorous statistical model.

Our primary goal at this point is to refine the clustering algorithm and to collate results from multiple reports in order to account for complexities we have seen in the data, particularly list bias. Also, we intend to collate additional reports from diverse geographical locations (such as Asian language reports) to broaden the scope of the phishing study. In this paper we have used capture-recapture separately within strata and across time; in the future we plan to study the effects of list bias, heterogeneity, trap effects and open populations simultaneously with a fully Bayesian statistical approach.

10. REFERENCES

[1] M. Abu Rajab, J. Zarfoss, F. Monrose, and A. Terzis. My botnet is bigger than yours (maybe, better than yours): why size estimates remain challenging. In *Proceedings of the first annual workshop on hot topics in botnets*, March 2007.

[2] L. Briand, K. Emam, B. Freimut, and O. Laitenberger. A comprehensive evaluation of capture-recatpure models for estimating software defect content. *IEEE Transcripts of Software Engineering*, 26:518–540, 2000.

[3] W. S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74:829–836, 1979.

[4] M. Collins, T. Shimeall, S. Faber, J. Janies, R. Weaver, M. De Shon, and J. Kadane. Using uncleanliness to predict future botnet addresses. In *Proceedings of the 2007 Internet Measurement Conference*, October 2007.

[5] E. Cooke, F. Jahanian, and D. McPherson. The zombie roundup: Understanding, detecting and disturbing botnets. In *Proceedings of the First Workshop on Steps to reducing unwanted traffic on the internet (SRUTI)*, July 2005.

[6] D. Dagon, C. Zou, and W. Lee. Modeling botnet propagation using time zones. In *Proceedings of the 13th Network and Distributed Security Symposium (NDSS)*, February 2006.

[7] J. Darroch, S. Fienberg, G. Glonek, and B. Junker. A three-sample multiple-recapture approach to census population estimation with heterogenous catchability. *Journal of the American Statistical Association*, 88:1137–1148, 1993.

[8] S. Fienberg. *The Analysis of Cross-Classified Categorical Data*. MIT Press, 1980.

[9] F. Freiling, T. Holz, and G. Wicherski. Botnet tracking: Exploring a root-cause methodology to prevent denial-of-service attacks. In *Proceedings of the 10th European Symposium on Research in Computer Security (ESORICS)*, September 2005.

[10] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. Technical Report 8, Soviet Physics Doklady, 1966.

[11] S. Lohr. *Sampling Design and Analysis*. Duxbury Press, 1999.

[12] P. McCullagh and J. Nelder. *Generalized Linear Models*. Chapman and Hall/CRC, 1989.

[13] D. Moore, C. Shannon, D. Brown, G. Voelker, and S. Savage. Inferring internet denial-of-service activity. *ACM Transactions on Computer Systems*, 24(2), 2006.

[14] T. Moore and R. Clayton. An empirical analysis of the current state of phishing attack and defence. In *Proceedings of the 2007 Workshop on the Economics of Information Security (WEIS)*, 2007.

[15] A. Ramachandran and N. Feamster. Understanding

the network-level behavior of spammers. In *SIGCOMM '06: Proceedings of the 2006 conference on Applications, technologies, architectures, and protocols for computer communications*, pages 291–302, New York, NY, USA, 2006. ACM Press.

[16] A. Ramachandran, N. Feamster, and D. Dagon. Revealing botnet membership using DNSBL counter-intelligence. In *Proceedings of the 2006 USENIX workshop on steps for reducing unwanted traffic on the internet (SRUTI)*, 2006.

[17] J. Rawlings, S. Pantula, , and D. Dickey. *Applied Regression Analysis*. Springer-Verlag, New York Inc., 1998.

[18] R. Thomas and J. Martin. The underground economy: Priceless. *Usenix ;login:*, 31(6), December 2006.

[19] J. Wittes. Applications of a multinomial capture-recapture model to epidemiological data. *Journal of the American Statistical Association*, 69:93–97, 1974.