A Continuous Time List Capture Model for Internet Threats.

Rhiannon Weaver*

Abstract

To study rapidly evolving populations of Internet threats under views from multiple watch lists, we propose a hierarchical Bayesian model we call Continuous-Time List Capture (CTLC). Methodologically, CTLC is related to survival analysis under competing risks, in which individuals under study admit as many survival curves as there are sources of watch list data. We suggest a Weibull model for the lifetime of a file from birth to appearance on a watch list, and we propose a Markov-Chain Monte Carlo method for simultaneous estimation of birth times for individuals, Weibull rate parameters for lists, and the effects of heterogeneity in behavior or traits among lists and individuals. We describe a population study of unique malware files under the CTLC framework, and present a preliminary simulation study as well as future work.

Key Words: Population estimation, Mark-recapture models, Hierarchical models, Network security, Risk analysis

1. Introduction

As the Internet grows larger and more complex, it is becoming harder to secure a system against every possible attack. It is increasingly important to have an understanding of the relative risk of threats and the associated costs of exploits. The idea of network security as risk management has been around for some time, as espoused by Blakely et al. (2001), but the foundations of principled risk analysis– security metrics and robust estimation methodology– remain elusive. As recently as April 2007, Geer's address to the Congressional Subcommittee on Emerging Threats, Cybersecurity and Science and Technology stressed the need for security metrics as a first priority. As an overall threat assessment, researchers are often interested in measuring the size of various malicious populations, such as malware files, machines in a collection of compromised hosts (botnet), or phishing sites.

A difficulty in measuring internet population sizes comes from the sheer speed and adaptability of Internet phenomena, in terms of both location and heterogeneity. With the advent of botnets, fast flux web hosting, cloud computing and proxy or peer-to-peer networks, the IP address is becoming a much more ephemeral measure of location for phenomena (Abu Rajab et al., 2007; Lemos, 2007). But blacklisting IP addresses is still often the first line of defense from attack. Also, as the Internet matures, the number of unique files of all types, including malicious files, grows steadily larger. This growth is exacerbated by hackers' use of self-changing, or *polymorphic*, code to defeat signature-based anti-virus software (Stepan, 2005; Grimes, 2007).

Not only does data consist of short-scale, temporal events, but often it is available only in fragmented views of the whole. Internet threats such as malicious files, open proxies and phishing sites, are reported publicly by third parties as temporal *watch lists* available for example through a daily or weekly web feed. Private sources such as volunteer reporters or honeypots can also be available for those interested in developing catalogs. Though Internet watch lists are not bound by spatial and geographical location in the same way that observers of physical phenomena are, the Internet is subject to shortest routes, Autonomous systems, and cultural and language diversity that can tend to localize views. This can

*CERT/SEI, 4600 5th Ave, Pittsburgh, PA 15213

all lead to watch lists providing smaller "telescopes" into the overall view of large-scale threats.

In this research approach we look at statistical analyses for temporal list data, with the following goals:

1. Leverage multiple network viewpoints to evaluate the scope of what is observed by providing a population estimate.

2. Evaluate data sources according to their coverage and their ability to discover hidden threats.

3. Track factors that make individuals (eg. IP addresses hosting phishing scams or open proxies; malicious files; bots) easier or harder to find.

Mark-recapture or multiple-recapture models (Lohr, 1999, ch 12) are often applied to population estimation and relative catchability comparisons using data from multiple lists. But much of the literature focuses on estimating population sizes under the conditions of no births or deaths occuring during the interval of observation (a *closed* population). Open populations are often considered under the context of ordered, repeated sampling, allowing for migrations only between sampling periods (Schwarz and Arnason, 1996; Muthukumarana et al., 2007; Dupuis and Schwarz, 2007).

Because of the speed of propagation of Internet phenomena, the scale over which a population can be assumed closed is usually short, relative to the scale at which inference is desired. Under this constraint, a more pertinent goal might be

4. Estimate the time from inception or "birth" of an individual until that individual appears on a watch list.

To address these goals, we propose a hierarchical Bayesian model we call Continuous-Time List Capture (CTLC). Methodologically, CTLC is related to survival analysis under competing risks (Beck, 1979), in which individuals under study admit as many survival curves as there are sources of watch list data. We suggest a Weibull model for the lifetime of an individual from birth to appearance on a watch list, and we propose a Markov-Chain Monte Carlo method for simultaneous estimation of birth times for individuals, Weibull rate parameters for lists, and the effects of heterogeneity in behavior or traits among lists and individuals. This is a preliminary report of continuing work.

Section 2 describes the components of the CTLC model. Section 3 discusses prior work. Section 4 formulates a population study for malware files within the CTLC framework. Section 5 presents a preliminary simulation study, and Section 6 discusses some future directions and caveats.

2. Continuous time list capture

2.1 Model background and notation

Suppose a set of J sources maintain watch lists of individuals related to a particular threat. Suppose a total of N individuals are observed during an interval $[t_0, t_M]$. We can associate with each individual i and each list j a variable W_{ij} defined as:

$$W_{ij} = \begin{cases} 1, & \text{individual } i \text{ is observed by list } j \\ 0, & \text{otherwise.} \end{cases}$$

In addition to W_{ij}, temporal information $T_{ij} \in [t_0, t_M]$ is collected that records the time at which individual i was observed by list j. The value T_{ij} is subject to right-censoring,

in that $T_{ij} = t_M$ when $W_{ij} = 0$. Note that individuals for whom $W_{ij} = 0$ for all j are unobserved in the study.

We can re-express the capture-recapture goals in this temporal framework. Each individual i has an associated birth time B_i, in which it is released "into the wild" and becomes a target for capture by various lists. In pure birth processes, a population estimate can be obtained by estimating the number of individuals existing in $[t_0, t_M]$ that remain unseen by all lists through t_M.

Heterogeneous catchability among lists or individuals can also be expressed in terms of time. A "wily trout" (Kadane et al., 1999) is an individual that takes longer to be seen than a conspicuous one. A good watch list finds individuals quickly with little dependence on other available lists in the study. We express heterogeneity among individuals by associating each individual i with a vector $\mathbf{X}_i$ of descriptive covariates. For example, a malicious file may be characterized by the operating system on which it can be deployed, the type of behavior it exhibits such as keylogging or scanning, or a descriptive "family" name assigned by an AV-vendor ("Virut", "Zlob", "Allaple", etc). Note that the covariates $\mathbf{X}_i$ are a way to link the temporal CTLC model with results from previous (non-temporal) descriptive analyses. The outputs of cluster analysis, feature analysis or other descriptive and exploratory analyses are used as inputs for the CTLC model, in the form of the vector $\mathbf{X}_i$ associated with each individual.

Similarly, we can associate each watch list j with a vector $\mathbf{Z}_j$ of descriptive covariates. These can be identifiers such as country of origin, size and scope of the operation, method of collection, and others. Another layer of heterogeneity can be a baseline catchability rate α_j for the list across all individuals. The covariates for lists and individuals, as well as the baseline list catchability rates, all provide information about the associated catch times T_{ij} and censoring variables W_{ij}. We use a Poisson process based on individual covariates to describe the birth process, and a Weibull process (one for each list) to describe the amount of time individuals "survive" before appearing on a list.

The values $\mathbf{X}_i$ and $\mathbf{Z}_j$ are used as covariates in a generalized linear model of "trait effects" as follows. Let $g(\mathbf{X}, \mathbf{Z})$ be a function that maps any pair $(\mathbf{X}_i, \mathbf{Z}_j)$ to the K-dimensional space $\{(0, 1)\}^K$. We can think of this as essentially asking K "yes or no" questions about individual i and list j that are answerable from their respective covariate vectors.

2.2 Formal model

Figure 1 shows the relationships between individuals, lists, catch times and censor variables as a directed graph. Round solid boxes represent unobserved parameters that are estimated based on observed values. Square solid boxes represent observed data values. Dotted solid boxes indicate replication by J, N, across both values, or across the dimension of the feature space K. The diamond box represents the deterministic relationship due to the link function for the GLM. Arrows between square and round boxes indicate a probabilistic relationship between the elements, which will be described in terms of a distribution function $p(\text{child}|\text{parents})$.

The top right corner of the graph describes the birth process. The function $\pi(\mathbf{X}_i, B_i)$ is a Poisson generating process of individuals, which is application-dependent. For example, for families of related individuals (for example, packed or polymorphic variants of one virus), the birth process should be defined in two steps:

- Birth of the first occurrence;
- Birth of subsequent occurrences given birth of the first occurrence

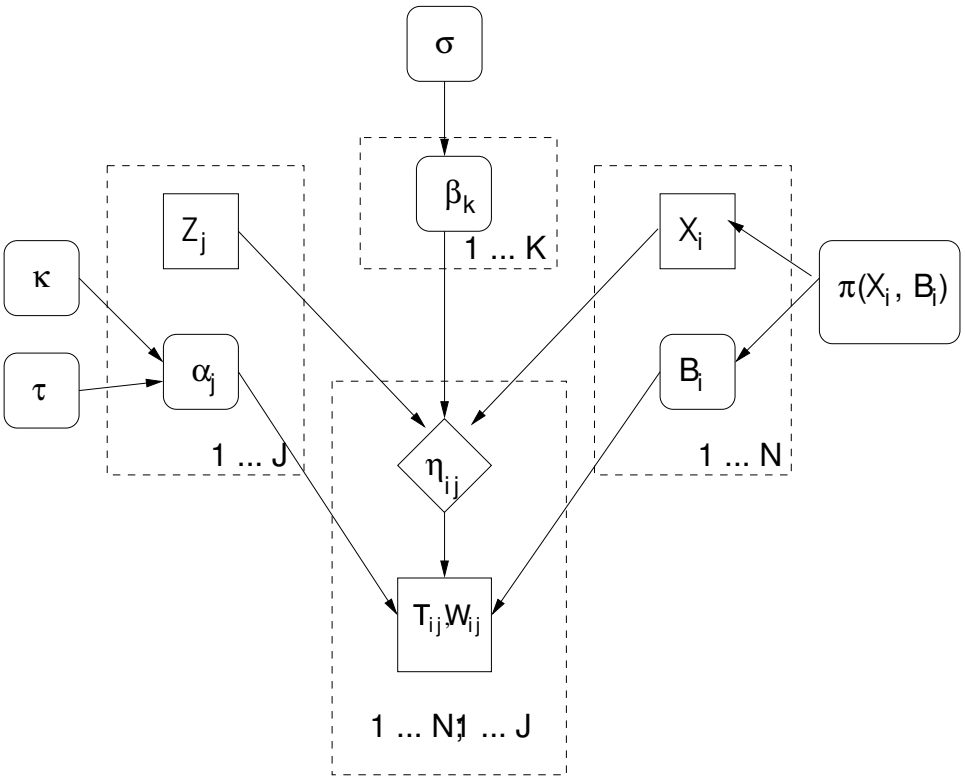

Figure 1: Graphical representation of the CTLC model. Round-edged boxes are unobserved parameters, and square boxes are observed data values. The diamond represents a deterministic relationship.

This specification allows for a "bursty" effect in births of similar individuals. The top left corner of the graph describes the characterization of watch lists, which are treated as fixed effects in the model.

The survival time of interest is the time from individual i's birth until capture by list j. In our notation, this lifetime is equal to:

$$L_{ij} = T_{ij} - B_i.$$

The linear predictor

$$\boldsymbol{\beta}' g(\mathbf{X}_i, \mathbf{Z}_j) = \sum_{g(\mathbf{X}_i, \mathbf{Z}_j)_k == 1} \beta_k$$

is the sum of trait effects for individual i and list j. Define the link function

$$\eta_{ij} = e^{-\boldsymbol{\beta}' g(\mathbf{X}_i, \mathbf{Z}_j)}.$$

The survival time of individual i on list j with baseline rate α_j is modeled as a Weibull(α_j, η_{ij}) variable:

$$p((T_{ij} - B_i)|\alpha_j, \eta_{ij}, B_i, W_{ij} = 1) = \frac{\alpha_j}{\eta_{ij}}(T_{ij} - B_i)^{\alpha_j - 1} e^{-\frac{(T_{ij} - B_i)^{\alpha_j}}{\eta_{ij}}}. \quad (1)$$

The trait effects β_k act to raise or lower the probability that an individual will survive past a certain time t before appearing on a list (see Figure 2).

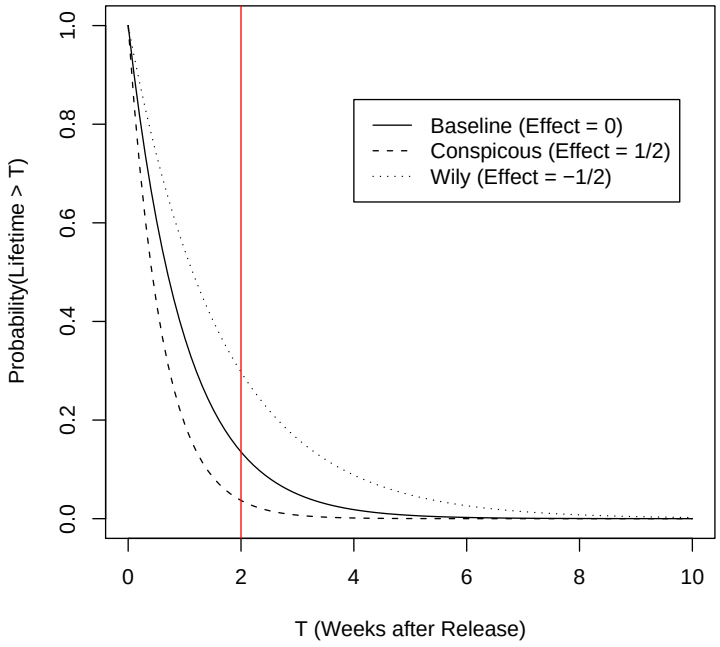

Figure 2: Trait effects acting on a Lifetime model.

The graph in Figure 1 is overly optimistic in considering all values (T_{ij}, W_{ij}) as observed, because the individuals for which $W_{ij} = 0$ for all j are in fact unobserved. The

likelihood for an individual i unobserved by list j through at least t_M is given by the cumulative probability

$$p((T_{ij} - B_i)|\alpha_j, \eta_{ij}, B_i, W_{ij} = 0) = e^{-\frac{(t_M - B_i)^{\alpha_j}}{\eta_{ij}}}$$

Thus, for each individual i and list j:

$$p((T_{ij} - B_i)|\alpha_j, \eta_{ij}, B_i, W_{ij}) = \left[\frac{\alpha_j}{\eta_{ij}}(T_{ij} - B_i)^{\alpha_j - 1} e^{-\frac{(T_{ij} - B_i)^{\alpha_j}}{\eta_{ij}}}\right]^{W_{ij}} \left[e^{-\frac{(t_M - B_i)^{\alpha_j}}{\eta_{ij}}}\right]^{1 - W_{ij}}$$

Suppose that individuals are grouped into one of a set of C "families" where it is presumed that multiple births are not independent. Let $\mathbf{X}_c$ be the characteristic vector associated with each subclass c, and let B_c be the first instance of a birth for subclass c.

We define the birth rate priors:

$$\begin{aligned}(B_i, \mathbf{X}_c | B_i = B_c) &\sim Pois(h_0(\mathbf{X}_c)) \\ (B_i, \mathbf{X}_c | B_i > B_c) &\sim Pois(h_1(\mathbf{X}_c, B_c)).\end{aligned}$$

Functions h_0 and h_1 are determined by the properties of individuals under study. Let σ, κ and τ be hyperparameters. We define the remaining priors:

$$\begin{aligned}\beta_k &\sim N(0, \sigma) \\ \alpha_j &\sim Gamma(\kappa, \tau).\end{aligned}$$

Let N_u be the set of individuals that are unobserved by any list. Let N_o be the set that are observed by at least one list. Let $\mathbf{X}_u$ be the matrix of covariates associated with all unobserved individuals. The product likleihood is described by:

$$L(\boldsymbol{B}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma, \kappa, \tau, N_u, \mathbf{X}_u | T, W, \mathbf{X}, \mathbf{Z}) =$$

$$\prod_{i=1}^{N_o} \prod_{j=1}^{J} \left\{ \left[\frac{\alpha_j}{\eta_{ij}}(T_{ij} - B_i)^{\alpha_j - 1} e^{-\frac{(T_{ij} - B_i)^{\alpha_j}}{\eta_{ij}}}\right]^{W_{ij}} \left[e^{-\frac{(t_M - B_i)^{\alpha_j}}{\eta_{ij}}}\right]^{1 - W_{ij}} \right\}$$
$$\times \left\{ \prod_{i=1}^{N_u} \prod_{j=1}^{J} \left[e^{-\frac{(t_M - B_i)^{\alpha_j}}{\eta_{ij}}}\right] \right\}$$

This formulation can be used in an MCMC framework for simultaneous parameter estimation.

3. Prior work: capture-recapture, survival analysis and generalized linear models

Capture-recapture population estimation is so named for its genesis in wildlife population studies. The simplest form, dating to Petersen (Petersen, 1896), requires repeated independent samples ($J = 2$) of a closed, homogenous population: no births or deaths occur between samplings, and the probability of sampling individual i is constant across all individuals. Table 1 shows an example. Capture profiles $\{W_{ij} : j = 1, 2\}$ for individuals are summarized as counts $m_{\mathbf{j}}$ in a 2×2 table, with the entry m_{00} equal to the (unknown) number of individuals not observed in either sample.

		In Sample 1?	
		No	Yes
In Sample 2?	No	m_{00}	m_{01}
	Yes	m_{10}	m_{11}

Table 1: Arrangement of counts for a 2-way simple capture-recapture study into a 2x2 table. The cell labeled m_{00} is the (unobserved) number of individuals that were not captured in either sample.

Peterson's estimate of the total population $n = N + m_{00}$ is equal to

$$\hat{n} = \frac{m_{+1}m_{1+}}{m_{11}},$$

where $+$ indicates a row or column summation.

The natural extension to multiple independent samples $(J > 2)$ for closed, homogenous populations followed through the early 20th century (Schnabel, 1938). For dependence among samples, Fienberg (Fienberg, 1972, 1980) introduced the use of the log-linear model, which was echoed by Cormack (1989). For J independent lists in a homogenous population, the probability of a binary capture profile $\mathbf{j} = (j_1, \cdots, j_J)$ does not depend on the individual i, and can be written as a product of capture probabilities associated with each list,

$$p(\mathbf{j}) = \prod_{\ell=1}^{J} \pi_{\ell j_\ell}$$

Aggregating across individuals, the expected cell counts $\mu_{\mathbf{j}}$ of the 2^J-way table of aggregated counts $m_{\mathbf{j}}$ can be written as a linear combination of parameter values on the log scale;

$$\log \mu_{\mathbf{j}} = \sum_{\ell=1}^{J} u_{\ell j_\ell}$$

This model can be seen as a multiplicative model of main effects. Accounting for list dependence equates to adding interaction terms into the log-linear model for the cell counts.

Applications to list data first arose in the mid-20th century in the context of human health studies, where researchers had access to records from multiple institutions for tracking individuals. Sekar and Deming (1949) used multiple lists to track birth and death rates as well as registration. Multiple ordered sampling methods such as dual-system estimation and post-enumeration survey have a history of use in the US Census dating back to the 1940s for example with Shapiro (1949).

Heterogeneity of individuals is typically modeled as covariate information in a general GLM framework when groups or distinguising features are known. Sekar and Deming demonstrate that population estimates based on aggregations across correlated sub-strata are biased toward an undercount of the true population. Seber (1982) provides general results describing bias as a function of correlation among capture probabilities of individuals, with the intuition that correlation of capture probabilities would generally be positive and would lead to consistent underestimation of populations, a particular form of Yule's association paradox (Yule, 1903). Kadane et al. (1999) formalize this argument by providing necessary and sufficient conditions under which this is the case.

In the unknown case, latent class or latent trait models are used. Darroch et al. (1993) present a multiple-recapture approach to heterogeneity of capture in the Census, using ordered repeated random samples to generate multiple views. They introduce the latent trait Rasch model (Rasch, 1980) to incorporate multiplicative heterogeneity in individuals and lists. The Rasch model describes the probability of capture, $\pi_{ij} = Pr(W_{ij} = 1)$ as a linear relationship on the logit scale:

$$\log\left[\frac{\pi_{ij}}{1-\pi_{ij}}\right] = \beta_j - \theta_i,$$

where θ_i is a random effect for individuals and β_j is a random effect for lists. Fienberg et al. (1999) explored the link between multiple list capture models, the Rasch model for heterogeneity of individuals, and a hierarchical Bayesian formulation of the Rasch model that explicitly calculates posterior distributions for individual parameters θ_i and list parameters β_j.

In networking and Internet characterization applications,Bradlow and Schmittlein (2000) apply a Bayesian model of heterogenous catchability in a closed population to explore the relative performance of six search engines, as well as to obtain population estimates of web pages characterized by key words. Briand et al. (2000) evaluate the use of capture-recapture models for estimating the number of errors or bugs in software applications.

Chan and Hamdi (2003) use capture-recapture methods to estimate the extent of total network resources in queue management schemes for routers. They apply both a homogenous capture-recapture model and a heterogenous model based on the Jackknife estimator (Burnham and Overton, 1978), that treats individual capture probabilities as nuisance parameters. In both cases, repeated samplings are modeled as independent draws over a closed population. Mane et al. (2005) use a capture-recapture method based on random walk sampling to estimate the number of nodes in a closed peer-to-peer network, assuming homogeneity among nodes, with motivations toward the study of open networks.

Extensions of capture-recapture methods to open populations generally focus on experiments where the sampling periods themselves are ordered in discrete time intervals, and assuming births, deaths and migrations occur between samplings, for example in Schwarz and Arnason (1996) and Dupuis and Schwarz (2007). But in the case of watch list data, all lists are active over the same time period, with births and deaths occurring throughout.

Defining population and catchability in terms of the time until visibility suggests a survival analytic approach, as opposed to the traditional multiple-recapture methodology. Each W_{ij} can be considered a binary indicator of observation in the multiple-recapture framework, but also as an indicator of right censoring in a survival model. In this framework, watch lists act in continuous time over an open and evolving population.

Parametric survival analysis can be considered a branch of generalized linear models in which the outcome of interest is a strictly positive distribution; traditional choices are exponential, Weibull or extreme value distributions (McCullagh and Nelder, 1989). Typically, non-temporal traits are modeled as multiplicative effects, estimated using partial likelihood or EM algorithms. An exponential model with unknown birth times and latent list parameters can be seen as a variant of the Rasch model using a complimentary log-log link as opposed to a logit link; Rasch also introduced a similar latent trait multiplicative poisson model with the logit model (Rasch, 1980). Cox (1972) introduced the widely-used nonparametric proportional hazards model, in which the survival curve is defined only where deaths occur.

In security analysis, Chen et al. (2006) use a survival model to quantify an economic analysis of the relative hazards of various vulnerabilities relative to the release of exploits;

they observe a "herding" behavior that indicates certain vulnerabilities are more attractive to hackers than others.

Analyses that allow for more than two states (alive or dead) were first proposed by Lagakos (1976) and later generalized to competing risks models by Beck (1979). Competing risks can be seen as a set of time-dependent covariates associated with each individual, that may increase or decrease the risk of capture. In the CTLC model, the capture or evasion of each watch list can be seen as a set of states that incur competing risks. Identifiability of regression parameters and states for competing risks models is discussed in Abbring and van den Berg (2003).

4. Applications to malware watch lists

The term *malware* is a portmanteau of "malicious software". Definitions vary among individuals and institutions, but a consensus is that a piece of malware is a program or file that damages or disrupts a computer system. Viruses, worms and Trojan horse programs are all malware. Malware can also refer to programs designed to corrupt or compromise a system for the gains of the designer; for example programs that install spyware, key loggers or back door controls (bots) to a system can also be considered as malware. Malware can propagate via standard network mechanisms such as email or through web browsing on the Internet. Malware files are collected by many different researchers, who maintain separate catalogues of files "captured" from the wild.

The number of unique malicious files found on web sites and on compromised hosts is increasing exponentially. In the past, analysts have catalogued these malicious files keyed by a unique hash of the bytes in the file. Signature-based detection from anti-virus (AV) vendors also relies on matching unique bit strings to known patterns. But new hacking trends are making organization by unique file difficult. For example, polymorphic viruses change as they replicate, and a single outbreak can result in thousands of unique individual files. Similar file structures do not translate to similar hash values, which makes it difficult to study trends in behavioral threats and attribution solely using hashes.

One way to analyze large-scale trends in the population is to categorize malware into families and variants based on behavioral traits. This imposes some structure and relationships on the ever-growing number of unique files, and engenders analytical questions such as:

- How much harder is it to find a key logger than a worm?
- Which watch list is the best at finding new browser exploits?
- What percentage of existing Allaple variants have all sources found?

This analysis requires two stages. A descriptive, non-temporal stage is required to label malware files with relavent family names and features, grouping them into related categories. Weaver and Sisk (2008) developed software tools for extracting these kinds of features as well as family names, from the names given to files by various anti-virus (AV) vendors, as well as methods for measuring agreement among vendors for family names. When a feature set is determined, the temporal stage of the analysis uses the CTLC model to make statements about the life cycle and catchability of these categorized files.

5. A preliminary simulation study with a simple survival model

Before formally addressing population estimates based on the CTLC model, the GLM and rate parameters were tested for basic identifiability using a simulation study. A simple

simulated population was devised using malware files as inspiration. An overdispersed Poisson distribution was used to simulate the births of 832 individual malware files from fifteen different families, over the course of 90 weeks. For the linear predictor, each family $c, c = 1, \cdots 15$, is associated with a baseline trait effect, $\beta_c \sim \text{Normal}(2.5, 1)$. Two malware traits were also specified independently for each file:

- Is the file described as a keylogger?
- Is the file described as a Trojan horse?

Keyloggers are programs that are designed to passively log user keystrokes, and to periodically update external servers. On the other hand, Trojan horse programs are designed to disable or damage computers once they have been downloaded and executed. This suggests that Trojan horse programs may be easier to find via user reports than keyloggers. To model these effects, the keylogger effect β_{16} was set to -0.5 as compared to the Trojan effect $\beta_{17} = 0.5$, encoding the fact that keyloggers are more "wily" than Trojans. Figure 3 displays the population of malware files used in the simulation study.

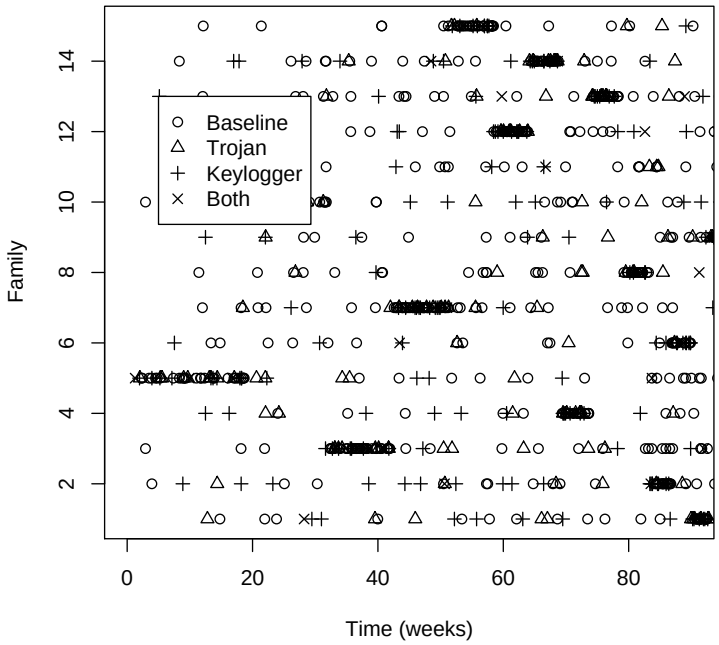

Figure 3: Simulated data for release of malware files from fifteen different families over the course of 90 weeks.

Three lists were simulated capturing files from this population, with capture probabilities independent across lists. The list baseline rates α_j were set equal to $1.5, 2.0$, and 2.5 for $j = 1, 2, 3$, respectively. Trait effects were assumed homogenous across lists, resulting in the basic additive linear predictor

$$\boldsymbol{\beta}' g(\mathbf{X}_i) = \sum_{c=1}^{15} \beta_c 1_{i \in c} + \beta_{16} 1_{\text{keylogger}} + \beta_{17} 1_{\text{Trojan}}.$$

List 1 Caught				List 1 uncaught			
		List 2				List 2	
		caught	uncaught		caught	uncaught	
List 3	caught	175	34	List 3	caught	30	36
	uncaught	40	47		uncaught	37	433

Table 2: Overall capture profile of 832 simulated malware files for three independent lists.

Table 2 summarizes the capture profiles of the three lists for the 832 simulated malware files.

As a simplifying assumption, birth times were assumed known for all captured individuals. In real-world data, some information on births may be extracted from analysis of malware files captured in the wild, but births will generally be represented as prior distributions possibly subject to hard constraints based on expert knowledge. Diffuse priors were used for α_j, while the true Normal$(3, 1)$ family prior was used as a prior for both family trait effects and the Keylogger and Trojan effects.

An MCMC chain using Metropolis-Hastings sampling with Normal or truncated Normal proposal steps was run for 3000 iterations to examine parameter estimation and identifiability for the CTLC model in this simple setting. Figures 4 and 5 show some graphical output from the chain for α_j and β_k. Although estimates for α_j appear to quickly reach stable values in the chain, the values appear to universally underestimate the three list rates. On the other hand, a plot of posterior means vs. true values for β_k shows a tendency to overestimate baseline family trait effects and the additive keylogger and Trojan effects. However, the difference $\beta_{17} - \beta_{16}$ yielded a posterior mean of 1.06, suggesting that while the baseline effects may suffer from identifiability issues or prior sensitivity, the relative trait effects are accurate. Future work is needed to determine the effects of this prior sensitivity on the ability of the model to produce accurate population estimates.

6. Discussion and future directions

In the statistical mark-recapture literature, multiple recaptures are generally modeled as the result of either multiple lists or multiple discrete sampling periods. Because Internet threats comprise short temporal events in open populations, and they are recorded through a lens of watch lists, population estimation models for these phenomena need to incorporate both of these sources of multiplicity. The Continuous Time List Capture (CTLC) model is a preliminary attempt to devise a formal framework for population estimation in this setting, using ideas from both mark-recapture models and survival analysis.

The CLTC model is still in its preliminary stages, both in model development and in applications for measuring populations of internet threats tracked by watch lists. Although the likelihood has been devised as a generative model for the birth and observation of malicious populations, preliminary analysis suggests that the model may suffer from some identifiability problems, and may also require subjective expert knowledge in order to produce trustworthy population estimates. Future work includes:

- Deriving formal population estimates from the likelihood and incorporating overall population estimation into the MCMC framework;
- Addressing the availability of birth data and the effect of unknown births on the ability to estimate other parameters (including populations) in the model;
- Addressing issues of model identifiability with increasing complexity in list dependency beyond independence;

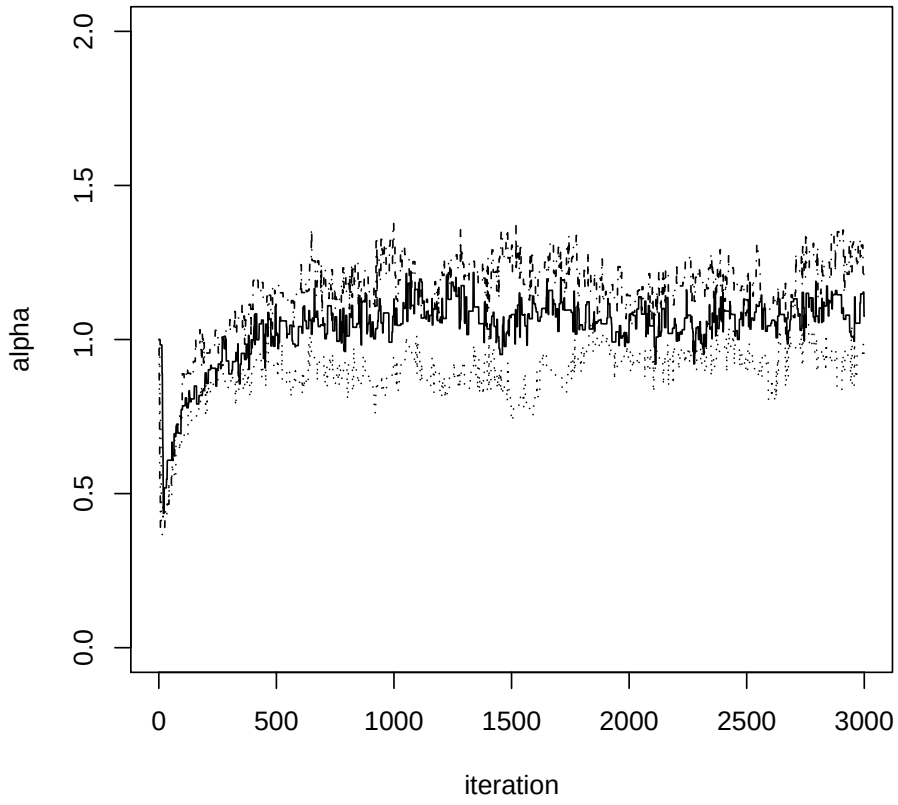

Figure 4: MCMC output (values vs. iteration) for estimation of list rates. Iterations of α_j show stable list rates with iterations in the chain but severe underestimation of parameters.

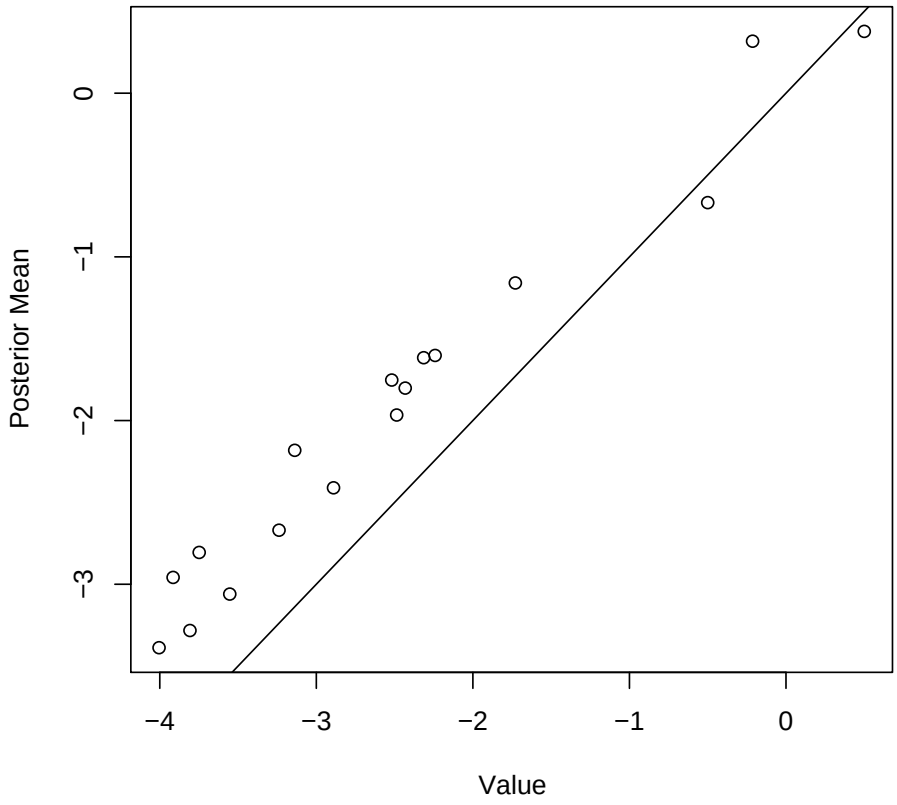

Figure 5: MCMC output (posterior mean vs. true value) for estimation of trait effects shows evidence of systematic overestimation of β_k.

- Incorporating measures of uncertainty into covariates (for example, conflicting family assignments of a file among vendors).

References

Abbring, J. and van den Berg, G. (2003). The identifiability of the mixed proportional hazards competing risks model. *Journal of the Royal Statistical Society: Series B*, 65(3):701–710.

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

Beck, G. (1979). Stochastic survival models with competing risks and covariates. *Biometrics*, 35(2):427–438.

Blakely, B., McDermott, E., and Geer, D. (2001). Information security is information risk management. In *Proceedings of the New Security Paradigm Workshop*.

Bradlow, E. and Schmittlein, D. (2000). The little engines that could: modeling the performance of world wide web search engines. *Marketing Science*, 19(1):43–62.

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

Burnham, K. and Overton, W. (1978). Estimation of the size of a closed population when capture probabilities vary among animals. *Biometriks*, 65:25–633.

Chan, M. and Hamdi, M. (2003). An active queue management scheme based on a capture-recapture model. *IEEE Journal on Selected Areas in Communications*, 21(4):572–583.

Chen, P., Kataria, G., and Krishnan, R. (2006). An economic analysis of the strategic interaction among comuter security attackers. In *Workshop on Information Systems and Economics (WISE)*.

Cormack, R. (1989). Log-linear models for capture-recapture. *Biometrics*, 45:395–413.

Cox, D. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society: Series B*, 34:187–220.

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

Dupuis, J. and Schwarz, C. (2007). A bayesian approach to the multistate jolly-seber capture-recapture model. *Biometrics*, 63:1015–1022.

Fienberg, S. (1972). Multiple-recapture census for closed populations and incomplete contingency tables. *Biometrika*, 59:591–603.

Fienberg, S. (1980). *The Analysis of Cross-Classified Categorical Data*. MIT Press.

Fienberg, S., Johnson, M., and Junker, B. (1999). Classical multilevel and bayesian approaches to population size estimation using multiple lists. *Journal of the Royal Statistical Society: Series A*, 162(3):383–405.

Grimes, R. (2007). Stopping malware that mutates on demand. http://www.infoworld.com/article/07/10/26/43OPsecadvise_1.html. InfoWorld online article.

Kadane, J., Meyer, M., and Tukey, J. (1999). Yule's association paradox and ignored stratum heterogeneity in capture-recapture studies. *Journal of the American Statistical Association*, 94(447):855–859.

Lagakos, S. (1976). A stochastic model for censored survival data in the presence of an auxiliary variable. *Biometrics*, 32:551–559.

Lemos, R. (2007). Fast flux foils bot-net takedown. http://www.securityfocus.com/news/11473. SecurityFocus online article.

Lohr, S. (1999). *Sampling Design and Analysis*. Duxbury Press.

Mane, S., Mopuru, S., Mehra, K., and Srivastava, J. (2005). Network size estimation in a peer-to-peer network. Technical Report TR 05-030, University of Minnesota Department of Computer Science and Engineering.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman and Hall/CRC.

Muthukumarana, S., Schwarz, C., and Swartz, T. (2007). Bayesian analysis of mark-recapture data with travel-time-dependent survivial probabilities. *The Canadian Journal of Statistics*, 35(2).

Petersen, C. (1896). The yearly immigration of young plaice into the limfjord from the german sea. *Report of the Danish Biological Station*, 6:5–84.

Rasch, G. (1960 (expanded 1980)). *Probabilistic models for some intelligence and attainment tests*. The University of Chicago Press.

Schnabel, Z. (1938). The estimation of the total fish population of a lake. *American Mathematical Monthly*, 45:348–352.

Schwarz, C. and Arnason, A. (1996). A general methodology for the analysis of capture-recapture experiments in open populations. *Biometrics*, 52(3):860–873.

Seber, G. (1982). *The estimation of animal abundance and related parameters*. London: Charles Griffen.

Sekar, C. and Deming, E. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association*, 44(245):101–115.

Shapiro, S. (1949). Estimating birth registration completeness. *Journal of the American Statistical Association*, 45:261–264.

Stepan, A. (2005). Defeating polymorphism: beyond emulation. http://downloads.microsoft.com. Microsoft Corporation white paper.

Weaver, R. and Sisk, M. (2008). A Trojan by any other name: analysis of malware naming conventions across vendors. *CERT Annual Research Report*, pages 13–16.

Yule, G. (1903). Notes on the theory of association of attributes in statistics. *Biometrika*, 2:121–134.