IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 19, 2020, accepted August 31, 2020, date of publication September 4, 2020, date of current version September 17, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3021760

Survival Analysis and Prediction Model of IP Address Assignment Duration

DAN KOMOSNY[1] AND SAEED UR REHMAN[2]
[1]Department of Telecommunications, Brno University of Technology, 616 00 Brno, Czech Republic
[2]College of Science and Engineering, Flinders University, Adelaide, SA 5042, Australia

Corresponding author: Dan Komosny (komosny@vut.cz)

ABSTRACT IP addresses of end hosts change when they are re-assigned. We apply survival analysis, which is commonly used in healthcare, on IP addresses to predict their assignment duration (their lifetime). We propose a survival parametric model based on a history of 6 years of address assignments on a worldwide scale. Our model outperforms alternative models both from short-term and long-term views. The custom modelling is also discussed as address assignment varies across Internet service providers (ISPs) and autonomous systems (ASs). A predictable address assignment duration has many applications, including source reputation, topology mapping, and geolocation. We describe a use-case in fraud prevention, where the proposed model is used as a trigger for two-factor authentication. The created dataset of addresses assignment durations is made publicly available.

INDEX TERMS IP address, survival, lifetime, host, assignment, security, IPv4, IPv6.

I. INTRODUCTION

Information about the IP address assignment duration is a key prerequisite for IP-to-host association, which is used in many applications, including geolocation, source reputation, topology mapping, and security. These applications work with the assumption that the association is valid for some time after the direct IP address observation, such as when a host accesses a service. However, addresses of end hosts change as being re-assigned, and the assignment duration is limited. A predictable duration of address assignments is therefore important for proper application implementations and future improvements.

In this work, we apply the survival analysis on IP addresses assignment duration (their lifetime). Such an approach, to the best of the authors' knowledge, has not been previously used. Survival analysis is typically used in healthcare to calculate the life expectancy of patients under specific observations (indicated disease, administered drug, underwent surgery, etc.). In our concept, the survival birth event is when a new IP address is assigned to a host and the death event when the address is changed, or the host becomes unavailable (based on a condition described later in Section IV).

The associate editor coordinating the review of this manuscript and approving it for publication was Lei Guo.

Based on the analysis, we predict the IP address lifetime by a parametric piecewise survival model. Our motivation for the model was that the best-alternative models used in healthcare or other areas did not accurately describe the lifetime of IP addresses. This is particularly shown by their accuracy comparison against a non-parametric survival estimator. The best-alternative models exhibit inaccurate address lifetime predictions, which our model aims to solve.

We define the piecewise model based on analysis of address assignments over 6 years on a worldwide scale, which we made publicly available at [1]. The model parameters are derived from the observed patterns in the survival hazard rate, which shows changes at specific days of address lifetime. We used these days for setting the piecewise model breakpoints. For these delimited pieces, we found the optimal value of the exponential survival. The model outperforms the best-alternative models.

We further discuss the custom modelling for local predictions as some ISPs or ASs use specific configurations for address assignments. This modelling may take advantage of the survival approach as the local datasets may not cover enough history of address assignments. Not enough historical records mean a more significant portion of addresses without the death event observed (i.e. they did not change until the end of observation). Excluding these right-censored addresses would lead to underestimation of the address life expectation.

Inclusion of their partial assignment durations would also produce wrong results.

Finally, as an application demonstration of the survival analysis, we elaborate a use-case in fraud prevention, where the model is used as a trigger for two-factor authentication.

The dataset used in this work is described in Sec. III. The IP address lifetime evaluated by the best-alternative parametric models is described in Sec. IV. The general piecewise exponential model is introduced and validated in Sec. VI. The custom modelling is covered in Sec. VII. The application use-case is given in Sec. VIII.

II. RELATED WORK

Paper [2] dealt with changes of IP addresses with a focus on the related events (reasons behind changes). The work was based on the RIPE Atlas [3]; specifically, the probe connection logs were analysed. The number of usable probes in the time of the paper (2015) was 3038, located in 156 countries. The time period investigated was 12 months. The relevant finding was that some ISPs re-assigned IP addresses periodically with large differences from 24 hours to weeks. IPv6 addresses were omitted from this work for two reasons. First, there was a low number of sole IPv6 hosts (237) and, second, the dual-stack hosts could not be used, as the method was based on analysing probe connections to the central controller. The two types of addresses often alternated at the dual-stack hosts and therefore the information about the duration of the same IPv4 address per probe was not obtainable (consecutive connections with IPv4 address were rare). Investigation of IPv6 address changes and comparison to IPv4 was left as future work.

Work [4] dealt with the relation of IP addresses and hosts with a focus on how ISPs organise their address space. Active measurements (interpolation of the timestamps of acknowledgement messages) were carried out to estimate how long the hosts kept the same IP address. The result was that the majority of addresses had the DHCP session duration in the range of 0–30 hours (approx.). The data were taken from the Shatel ISP with about 750,000 DHCP session logs. The mean number of addresses per user within 24 hours was around 5. The median address session duration was found for four additional ISPs – AT&T 40 hours, British Telecom 9 hours, Deutsche Telecom 8 hours, and Orange 7 hours. The authors mentioned a possibility of address classification according to their usage in order to obtain more precise results. The address categories mentioned were mobile, wireless, home, small and large businesses. The WHOIS database was suggested as a source of data for the address classification. This database was used for finding addresses associated with retail/business, content distribution network, and infrastructure by specific IP address domains (e.g. t-ipconnect.de).

Paper [5] studied the possibility of user tracking by their device IP addresses. For this purpose, the authors evaluated the stability of the addresses assigned to the end devices (or NAT device). The addresses were collected via two web browser extensions, which reported the device address every four hours. The address duration was defined as the time from the first to the last same address occurrence (actually, the device may have used other addresses between the first and last same address). The authors called this duration as the retention period. The true reason for the address change was not known as the change could have happened in the same network or the device may have moved to a new network. Around 2,000 users reported approx 35,000 unique addresses that were evaluated. The evaluation period was 111 days. The result was that 87 % of users retained at least one address for more than one month. The mean address retention period was approx. 9 days and 11 % of them had the period longer than a month. The authors also evaluated a hypothesis that addresses may be used for user identification. The Jaccard similarity was calculated between the user address sets. The result was reported as 93 % of the users had a unique address set.

Paper [6] dealt with spam reduction by identifying the common properties of spam bots. The authors' finding was that addresses (IP address blocks with a given prefix) of spamming hosts frequently changed as only 42 % of them had a usage duration longer than 14 hours. On the other hand, 70 % of the non-spamming host address prefixes had a duration longer than 14 hours. For a duration longer than 28 hours, the percentages of the same address prefixes decreased to 22 % for spamming hosts and 44 % for non-spamming hosts. The uptime value was defined as the median of seconds of all periods when an address was active. The maximum was 226 hours, which was the time span of the evaluation. It was concluded that the use of IP blacklists might cause problems by blocking legitimate hosts with an address that was previously used by a spamming host (previously identified as a spammer).

Work [7] focused on spam botnets and blacklists. Spam bots were tracked by their IP addresses. The authors' observation was consequently applied to mail delivery logs to identify other hosts that have similar behaviour to spam bots. Several botnets were used in the evaluation, including Rustock and MegaID. The analysed time span was from September 2010 to February 2011. The found fraction of static to dynamic addresses was 15 % for Rustock and 4 % for MegaID. An address was considered as dynamic if it was spotted only once during the evaluation period. If the same address was observed multiple times, it was considered as static. It may be indirectly concluded that about 4-15 % of spam bots addresses per host were unchanged over a period of 6 months.

Paper [8] dealt with a prediction of the next assigned address. The prediction was processed from the attacker viewpoint. The addresses studied came from two major cloud service providers – Amazon Web Services (AWS) and Google Cloud Platform (GCP). Different strategies were used to collect the addresses, such as working with geographical regions. Around 314,000 addresses were collected from AWS, of which 89,000 were unique. Around 29,000 addresses were collected from GCP, of which 42 were unique. The collection period varied per platform and per area

TABLE 1. Overview of related work in terms of outcome type, application, and novel idea.

Work	IP ver.	Node type	Main outcome	Application/use-case	Novel idea[2]
[2]	v4	End node	Analysis	-	Power outages vs. addresses
[4]	v4	End node	Analysis	-	-
[5]	v4	End node	Analysis	On-line privacy protection	Similarity of address sets
[6]	v4	Server	Analysis	Spam reduction	-
[7]	v4[1]	Server	Spam bot tracking	Spam reduction	Characterisation of addresses
[8]	v4	Server	Analysis; Next address prediction	DoS attack prevention	Formal description of DoS
this	v4, v6	End node	Analysis; Address lifetime prediction	Cyber fraud prevention	Survival of addresses

Specific results are given in related work description.
[1] Not stated; various datasets were used.
[2] By our opinion.

from 42 to 109 days. The authors evaluated the time in days an attacker would need to collect the address prefixes (first three bytes) for a reasonable prediction of the next assigned address. The result was 54 days (max) for AWS and 39 days for GCP. For the prediction feasibility evaluation, 70 % of the collected datasets were used as the machine training data, the remainder were used as the test data. The address prediction (three bytes) was successful at least at 90 % with one exception. The results suggested that an attacker may be able to correctly predict the next allocated addresses to perform DoS (Denial of service) attack.

IP address-to-host association has many other applications. As an example of novel use, work [9] deals with the identification of web pages that a host accessed just by observing the destination addresses from the communication. When accessing a web page, many objects are loaded from subresources at different servers (e.g. images). All these servers are visible by their address. This set of addresses forms a page-load addresses fingerprint. Such fingerprint may identify a page accessed even if the same address is shared by different sites. The dataset used in [9] consisted of one million sites. The authors' finding was that more than 95 % websites had a unique page-load destination address fingerprint. This allows page identification, provided that a database of address fingerprints is available. The question of address changes over time was not elaborated.

Some commercial organisations, especially dealing with IP targeted marketing, claim that addresses change very rarely, such as [10] and [11], state that addresses were kept for the same households for seven/nine months. On the other way, there are initiatives to change IP addresses frequently for improved privacy [12]. There are also methods to detect IP changes at intermediate devices – paper [13] describes a method to detect a change of WiFi AP addresses. It is based on authenticated web requests from the AP clients. The first request of the procedure, which is authenticated by a cookie, is from the address of the AP. If the consequent request within a short time is from a different address, the change is detected. Other related papers [14] and [15] dealt with optimal DHCP configuration. Other papers [16] and [17] dealt with lists of addresses that are persistent over time, for example, to facilitate long-term global measurements.

Finally, a summarised overview of the most significant related work is given in Table 1. The table compares the related work in terms of the address version studied, whether the addresses in focus were assigned to servers or end nodes, the main outcome type, the intended application or use-case, and the novel idea presented. The related information about this work is shown in the last row.

III. DATASET OF IP ADDRESS ASSIGNMENTS

The dataset of IP address assignment durations was processed from the RIPE Atlas data [18]. The Atlas is a set of measurement probes [3] installed by users in their networks. The probe IP address along with other information is archived and publicly available. The probes can be either dedicated hardware boxes or software application. The hardware probes are distributed by RIPE NCC upon a request. The aim is to distribute the probes evenly across autonomous systems and geographical regions. The users may also run the probe software on their devices. In this case, the probe installation is not restricted. The software is freely downloadable as source code [19] or as a platform-specific build, including versions for CentOS, Debian, and Docker. The probe owners provide a description of the environment where the probe is installed, such as 'Fibre', 'Academic', 'No NAT', and 'dual-stack'. They are also responsible for their maintenance, though many probes alternate between operational and non-operational states and some of them are permanently abandoned.

The IP address is assigned to a probe the same way as to other devices in the network (i.e. desktop, laptops, etc.). We worked only with the probes installed behind NAT (probe installation environment was described as 'NAT'), thus representing the end hosts. The probe address along with its operational status is provided by the Atlas in daily archive snapshots. We processed these archive snapshots for the addresses of the operational probes only; addresses of non-functional probes were excluded. The processing covered parsing each daily snapshot and merging the obtained addresses for a particular probe. The created dataset, suitable for address survival analysis, starts on 13/3/2014 and ends by 26/6/2020, covering 2,288 days (one day snapshot 21/02/2018 was corrupted). The geographical distribution of the active addresses from the latest daily snapshot is shown

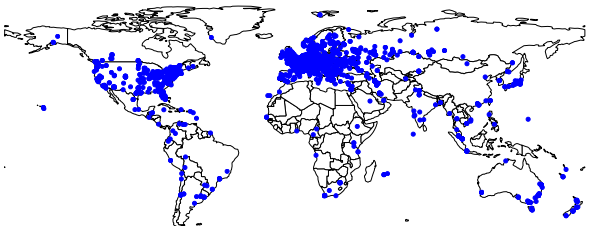

FIGURE 1. Geographical distribution of active IPv4 addresses from the latest daily snapshot; most addresses come from Europe and the USA.

in Figure 1; the majority of addresses came from Europe and the USA. The total number of address records was approx. 626,000 for IPv4 and 195,000 for IPv6. The created dataset is made publicly available at [1] for further exploration.

IV. IP ADDRESS LIFE EXPECTANCY

In this section, we analyse the IP address assignment durations. In terms of the survival analysis, the birth event is defined as the day when a new address started to be used by a host, the death event as the day of its change. If the death event was not observed (the address was assigned to a host in the most current day of the dataset), the death-observation tag was set as 'False'; otherwise as 'True'. The IP addresses with this tag set were right-censored in the survival analysis. Further, the address death event was also triggered if the host became unavailable. This was when the probe status changed to non-operational. If the probe later returned to the operational state, the used IP address was born and included in the analysis. An example is shown in Figure 2. The listing shows the survival data for two Atlas probes with these fields: probe ID, active address, duration – number of days of address assignment, and death tag. The address changes were observed in day resolution. For example, a change between the 1st and 2nd day is indicated by an assignment duration of 1.5 days. In the example, three addresses were assigned to the first probe. The first two addresses were changed, the death event was observed, and the tag was set as 'True'. The third address of the first probe was still assigned at the end of the observation, and the death event was not observed and the tag was set as 'False'. The second probe used two addresses. Both addresses were changed during the observation. At the end of the observation, the probe was not operational, and therefore there was no address with the death tag set as 'False'.

ID	IP address	Days	Dead
1000	144.134.101.42	4.5	True
1000	144.134.219.219	1.5	True
1000	144.134.112.143	13.5	False
1000019	112.133.206.18	10.5	True
1000019	112.133.244.24	0.5	True

FIGURE 2. Sample of IP address assignment observations; full data at [1].

The address death observation is also graphically demonstrated in Figure 3. A sample of IPv6 addresses is shown

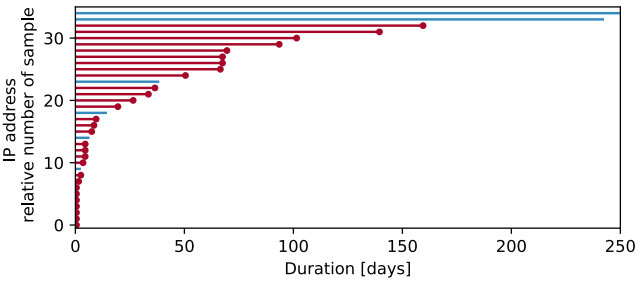

FIGURE 3. Example of right-censoring in address lifetime analysis (sample). Red lines – death event was observed (address changed). Blue lines – death was not observed (address was not changed up to the last day of observation).

for demonstration clarity. The blue lines show the address assignment durations in days that did not end (death tag is set as 'False'). The red lines shows address durations that ended – their change was observed (death tag set as 'True').

The survival function of address lifetime $S(t) = P(T > t)$ gives the probability of address surviving past t. T is the random lifetime and t is the time for which the death event was not observed. The function for the dataset was obtained by the Kaplan-Meier non-parametric estimator, which is

$$\hat{S}(t) = \prod_{t_i \leq t} \frac{n_i - d_i}{n_i}, \tag{1}$$

where t_i is the time when at least one address per host was changed (death was observed), d_i is the number of addresses that changed at t_i, and n_i is the number of addresses that were assigned (alive/at risk of death, also censored) up to the time t_i.

The use of the Kaplan-Meier estimator is based on fulfilling three assumptions about the input data [20]. i) "At any time patients who are censored have the same survival prospects as those who continue to be followed.", ii) "The survival probabilities are the same for subjects recruited early and late in the study" and iii) "The event happens at the time specified". For our data, the first assumption is met as there is no difference in censored and non-censored addresses in terms of their survival, i.e. censoring is independent of the likelihood of address change. The second assumption is met as there is no difference in terms of survival of addresses that appeared early or late in the evaluation time span (we assume that at the global scale, the address assignment duration is homogeneous in time during six years). The third assumption is met as the address changes are observed daily, i.e. in a day resolution, which is a short interval relative to the whole evaluation time span (this assumption may otherwise cause problems in healthcare when the exact date of the event is not known as only evaluated when patients are examined, which may be in long and irregular intervals.).

The survival probability was evaluated for the address version. The resulting survival curves are plotted in Figure 4.

The survival curves have a long tail of low-probability values up to approx. 2000 days. We attempted to fit the curves

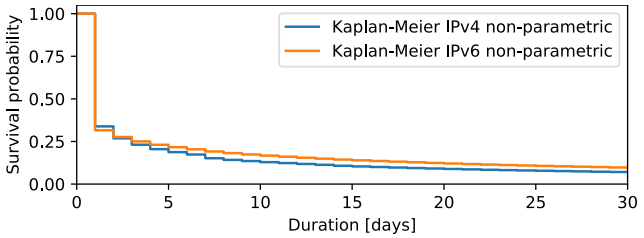

FIGURE 4. Probability of IP address survival. Long lifetime values to approx. 2000 days are omitted for clarity.

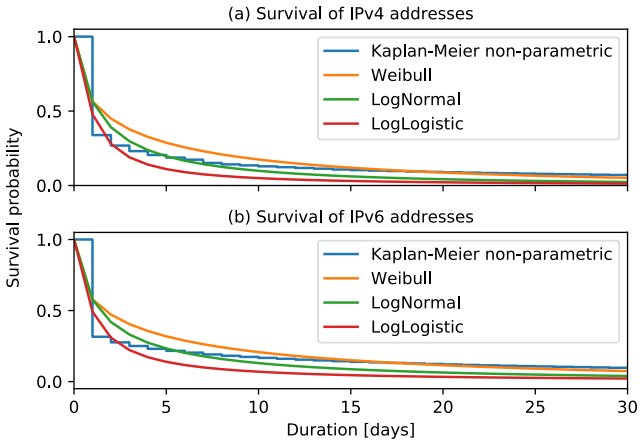

FIGURE 5. Survival curves fitting for (a) v4 and (b) v6 addresses by the common models. None of the models shows a good match.

by the common survival parametric models. The result for both v4 and v6 addresses is shown in Figure 5. The curves show that there was not a good approximation by any of the best-alternative models, which were Weibull, LogNormal, and LogLogistic.

To accurately assess the goodness of fit of the models, we used two techniques: i) calculation of survival probability difference at specific days and ii) survival curve similarity over a period of days. The first assessment is used for the short-term comparison, as these are the biggest discrepancies in the survival functions. According to the observed errors in Figure 5, we set these days as $\{2, 5, 10\}$. The second assessment compares survival curves up to a number of days by the restricted mean survival time (RMST), which is

$$\mathrm{RMST}(t) = \int_0^t S(u)\mathrm{d}u. \tag{2}$$

RMST calculates the area under the survival curve up to the time t. For a good fit, the difference in RMST (i.e. the area between the survival curves) should be minimal. This assessment compares the survival curves from the long-term view, and we set the restrictions t as $\{30, 60, 365\}$ days. Sample difference in RMST with a time restriction $t = 30$ is shown in Figure 6. It shows the graphically-compared best model for the 5$^{\text{th}}$ day, which is LogNormal. The difference in RMST (the delimited area between the curves) was approx. 1.02. The figure also shows that the short and long-term results are

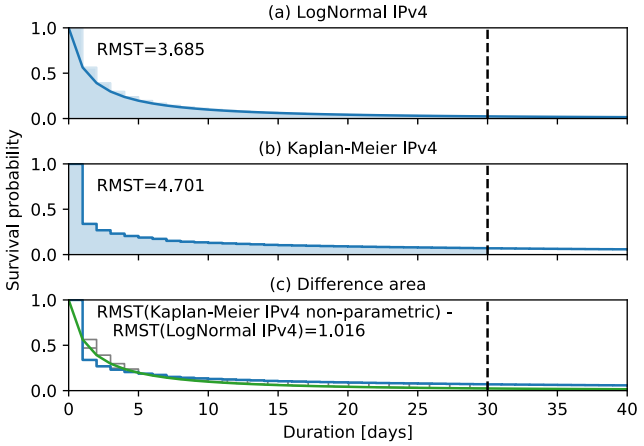

FIGURE 6. Long-term comparison of goodness of fit of (a) the LogNormal parametric model ($\mu = 0.26$ $\sigma = 1.59$) against (b) the Kaplan-Meier non-parametric estimator. Compared is (c) the restricted mean survival time (RMST) with a time restriction $t = 30$.

not consistent and therefore support the short and long term separate evaluation.

Table 2 gives the comparison results for the best-alternative models by difference to the Kaplan-Meier non-parametric estimator. The model best-fit parameters are listed in brackets.

TABLE 2. Goodness of fit of the parametric models for v4 and v6 addresses lifetime. Compared is the survival probability difference at days and restricted mean survival time (RMST) up to days. Model best-fit parameters are in brackets.

Short-term – survival at day	2	5	10
Kaplan-Meier (non-par. estimator) **IPv4**	0.27	0.19	0.13
Weibull ($\lambda = 3.17$, $\rho = 0.48$) diff.	-0.18	-0.1	-0.05
LogNormal ($\mu = 0.26$, $\sigma = 1.59$) diff.	-0.12	-0.01	0.03
LogLogistic ($\alpha = 0.93$, $\beta = 1.25$) diff.	-0.01	0.08	0.08
Kaplan-Meier (non-par. estimator) **IPv6**	0.28	0.22	0.17
Weibull ($\lambda = 3.73$, $\rho = 0.46$) diff.	-0.2	-0.1	-0.04
LogNormal ($\mu = 0.34$, $\sigma = 1.75$) diff.	-0.14	-0.02	0.04
LogLogistic ($\alpha = 0.97$, $\beta = 1.11$) diff.	-0.03	0.08	0.1
Long-term – RMST up to day	30	60	365
Kaplan-Meier (non-par. estimator) **IPv4**	4.7	6.33	10.94
Weibull ($\lambda = 3.17$, $\rho = 0.48$) diff.	-0.57	0.19	4.22
LogNormal ($\mu = 0.26$, $\sigma = 1.59$) diff.	1.02	2.24	6.44
LogLogistic ($\alpha = 0.93$, $\beta = 1.25$) diff.	2.26	3.64	7.76
Kaplan-Meier (non-par. estimator) **IPv6**	5.56	7.78	13.8
Weibull ($\lambda = 3.73$, $\rho = 0.46$) diff.	-0.55	0.29	4.97
LogNormal ($\mu = 0.34$, $\sigma = 1.75$) diff.	1.13	2.61	7.59
LogLogistic ($\alpha = 0.97$, $\beta = 1.11$) diff.	2.63	4.42	9.46

V. SURVIVAL FUNCTION ANALYSIS

In order to obtain a better fit, we analysed the survival function properties. The function can also be expressed by its hazard $h(t)$, which gives the rate of address changes (deaths) during a time interval $(t, t + \mathrm{d}t]$ provided that the changes

have not occurred up to the time t

$$h(t) = \lim_{\mathrm{d}t \to 0} \frac{P(t < T \leq t + \mathrm{d}t | T > t)}{\mathrm{d}t}. \tag{3}$$

The integral of the hazard function gives the cumulative hazard

$$H(t) = \int_0^t h(z)\mathrm{d}z. \tag{4}$$

The cumulative hazard was obtained by the Nelson-Aalen non-parametric estimator, which is

$$\hat{H}(t) = \sum_{t_i \leq t} \frac{d_i}{n_i}, \tag{5}$$

where t_i is the time when at least one address per host was changed (died), d_i is the number of addresses changed at t_i, and n_i is the number of addresses assigned (survived/being at risk of death) up to time t_i.

The hazard rate $\frac{d}{dt}H(t)$ for v4 and v6 addresses is shown in Figure 7. The hazard first significantly increases as many addresses do not survive the first day, then it has a decreasing trend with some exceptions.

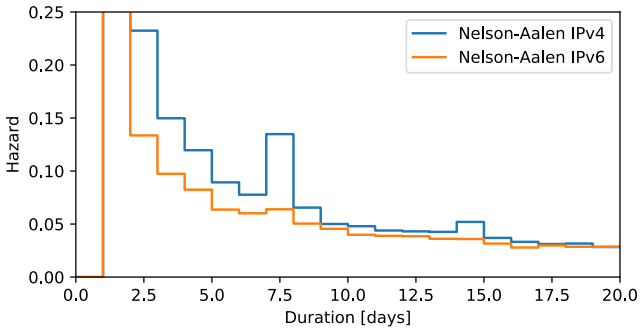

FIGURE 7. Hazard rate for v4 and v6 addresses survival. Hazard is limited to 0.25 for clarity.

VI. PIECEWISE PARAMETRIC MODEL

The observed hazard rate in Figure 7 suggests the application of the piecewise exponential model. This model is a set of exponential models between M breakpoints $\mathrm{T} = \{\tau_0, \ldots, \tau_{M-1}\}$. The cumulative distribution function $F(t)$ of the single exponential model is

$$F(t) = 1 - e^{-\lambda t}, \tag{6}$$

with its survival function

$$S(t) = 1 - F(t) = 1 - (1 - e^{-\lambda t}) = \mathrm{e}^{-\lambda t}. \tag{7}$$

The hazard function is $h(t) = \lambda$ and the cumulative hazard is $H(t) = t\lambda$. The hazard rate $h(t)$ for the intervals of the constant λ between M breakpoints T is

$$h(t) = \begin{cases} \lambda_0, & t \leq \tau_0 \\ \lambda_1, & t \in (\tau_0, \tau_1] \\ \cdots & \\ \lambda_M, & \tau_{M-1} < t, \end{cases} \quad t \in (0, \infty). \tag{8}$$

For the setting of λ values, we consider the input dataset properties. Its time span of 6 years is far beyond the scope of the target applications. Also, the probability of address survival beyond one year is low – $S(365) \doteq 0.005$ for v4 and $S(365) \doteq 0.006$ for v6. We, therefore, restrict the survival function $S(t)$ for $R = \{t \in \mathbb{N} | t \leq 365\}$ as $S\big|_R(t)$ (defined for all t in R). The observed hazard changes in Figure 7 suggest settings of the breakpoints at the 1st and 7th day. We set another point on the 30th day to cope with the long-tail values of the survival function. Additional breakpoints may be set for a better fit. We set the three points for clarity of presentation of the modelling process. By assuming the breakpoints $\mathrm{T} = \{1, 7, 30\}$ and the survival domain R, the cumulative hazard $H(t) = \sum_{i=1}^{t} h(i)$ is

$$H(t) = \begin{cases} \lambda_0, & t = 1 \\ (t-1)\lambda_1 + H(1), & t \in [2, 6] \\ (t-6)\lambda_2 + H(6), & t \in [7, 29] \\ (t-29)\lambda_3 + H(29), & t \geq 30, \end{cases} \quad \text{and } t \in R, \tag{9}$$

where the λ values are set for v4 and v6 addresses according to Table 3. The values were rounded for the clarity of calculation except for λ_0, which is important for the cumulative hazard offset.

TABLE 3. Value of λ between breakpoints T.

Hazard rate	λ_0^{-1}	λ_1^{-1}	λ_2^{-1}	λ_3^{-1}
IPv4	$\frac{25}{27}$	8	24	120
IPv6	$\frac{20}{23}$	13	29	110

Finally, the restricted survival function of the model is

$$S\big|_R(t) = e^{-H(t)}, \tag{10}$$

where $H(t)$ is the cumulative hazard defined in Eq. 9 with λ values provided in Table 3. The model fit is graphically shown in Figure 8 and the comparison numbers are shown in Table 4. The values are presented using the same techniques as in Table 2. The piecewise model shows the lowest difference values for both v4 and v6 addresses when compared to the best-alternative model, which is LogNormal. The largest improvement in survival prediction in the short-term range was at the 2nd day for v6 addresses. For this day, the probability difference of the best-alternative LogNormal model to the non-parametric Kaplan-Meier estimator was 14 %. The Piecewise exponential model had a difference in survival probability only of 2 %. In the case of the long-term evaluation, the largest improvement was on the 365th day again for v6 addresses. The alternative LogNormal model had an area-under-curve difference to the Kaplan-Meier estimator of 7.6, whereas the Piecewise exponential model had an area difference only approx. 1. We note that adding more pieces to the Exponential model would result in an even better fit and thus providing larger improvements in the survival

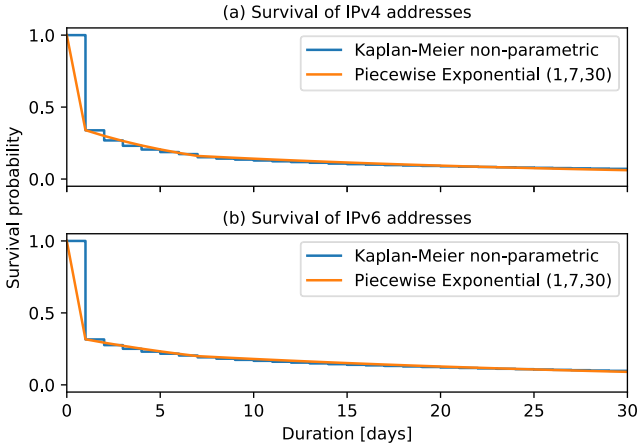

FIGURE 8. Piecewise exponential model goodness of fit with breakpoints (change in λ) T = {1, 7, 30} days. The model shows a good match for both (a) v4 and (b) v6 addresses.

TABLE 4. Goodness of fit of the piecewise exponential model derived from hazard rate analysis. Error (difference) values compared to the best-alternative models.

Short-term – survival at day	2	5	10
Kaplan-Meier (non-par. estimator) **IPv4**	0.27	0.19	0.13
Best alternative – LogNormal diff.	-0.12	-0.01	0.03
Analysis model – Piecewise exp. diff.	-0.03	-0.02	-0.01
Kaplan-Meier (non-par. estimator) **IPv6**	0.28	0.22	0.17
Best alternative – LogNormal diff.	-0.14	-0.02	0.04
Analysis model – Piecewise exp. diff.	0.02	-0.02	-0.01
Long-term – RMST up to day	30	60	365
Kaplan-Meier (non-par. estimator) **IPv4**	4.7	6.33	10.94
Best alternative – LogNormal diff.	1.02	2.24	6.44
Analysis model – Piecewise exp. diff.	0.28	0.28	-0.41
Kaplan-Meier (non-par. estimator) **IPv6**	5.56	7.78	13.8
Best alternative – LogNormal diff.	1.13	2.61	7.59
Analysis model – Piecewise exp. diff.	0.27	0.12	-0.94

probability modelling. We used the three pieces for clarity of presentation.

VII. CUSTOM MODELLING

The proposed general model was derived from the dataset with a long history of worldwide address assignment durations. However, assignment durations vary across ISPs (Internet service provider) and ASs (autonomous system). These differences are given by the custom configuration of the address leasing devices and, also, by various environmental/administrative factors, such as planned periodic changes, networking outages, device reconfigurations, and power stability. Specifically, work [2] studied the correlation of address changes to outages. The finding was that the likelihood of address change caused by outages varies across ASs. The other relevant finding was that the outage length affects the number of changed addresses.

For these reasons, we discuss the process of custom modelling for specific ISPs and ASs. A prerequisite for a custom model is the availability of a dataset of prior address assignments. Such custom dataset may not cover a long assignation history, as our dataset used for the general model. A shorter history means a larger proportion of addresses for which the change (death event) has not been observed yet.

The survival analysis copes with these unfinished observations, as demonstrated in Table 5 for our dataset. It shows that the censoring numbers are low compared to the total number of observed changes at each time interval, thus having a relatively small effect on the result. However, we deliberately used the survival analysis concept as the censoring allows extensions for custom modelling.

TABLE 5. Head of the life table for the dataset with a long history. The censoring is important for custom modelling based on limited datasets.

Event	Removed	Observed	Censored	Entrance	At risk
0.0	0	0	0	626532	626532
0.5	414425	414425	0	0	626532
1.5	44071	43987	84	0	212107
2.5	23472	23375	97	0	168036
3.5	16349	16301	48	0	144564

For the custom models, the intervals of constant λ will vary. Their setting is based on observation of the hazard rate $h(t)$, which is a derivation of the cumulative hazard $H(t)$. $H(t)$ is directly obtained for a dataset by Eq. 5 (Nelson-Aalen non-parametric estimator). We suggest the breakpoints placement at the points of hazard rate where the curve goes opposite to its general trend, as used in the general model. This process may be automated by software for batch modelling. However, care should be taken when obtaining $h(t)$ by derivation of $H(t)$. Some derivation software use by default a kernel smoother with bandwidth to smooth the resulting curve. Improper smoothing may produce hazard with the changes exaggerated or lost. For example, the Lifelines software [21] uses the Epanechnikov kernel

$$D(t) = \begin{cases} \frac{3}{4}(1-t^2), & |t| < 1 \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

Other way is to perform simple numerical derivation $\frac{H(t+\Delta t)-H(t)}{\Delta t}$ with the address assignment observation period Δt equal to a day. A suitable derivation of $H(t)$ for the breakpoints settings is shown in Figure 9. The linear approximation of the hazard between the breakpoints determines the λ values to be used in Eq. 9.

VIII. SECURITY USE-CASE

We give an example of use in security, particularly fraud prevention. Assume a user has been logging to an on-line service from the same device's IP address in the period of 10 days (up to now). What is the probability that the user will login after two days (during the 3rd day) from the same device using this IP address? (i.e. also from the same NAT network). The example is calculated for IPv4. We employ the conditional

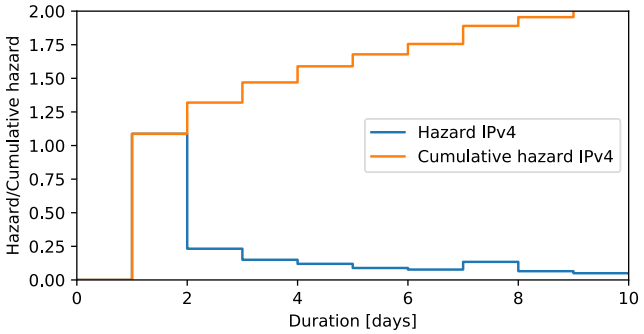

FIGURE 9. Derivation of hazard rate from cumulative hazard. Improper smoothing bandwidth may result in wrong breakpoints settings.

survival function

$$S(t|T > D), \tag{12}$$

where D is the number of days when the address was used by a host. Considering $t > D$ ($T > t \subset T > D$) and by $S(t) = P(T > t)$ we obtain

$$S(t|T > D) = \frac{P(T > t \cap T > D)}{P(T > D)} = \frac{P(T > t)}{P(T > D)} = \frac{S(t)}{S(D)} = \frac{e^{-H(t)}}{e^{-H(D)}}. \tag{13}$$

By substituting the hazard rates λ given in Table 3 to Eq. 9 we specify the cumulative hazards $H(t)$ and $H(D)$. The restricted survival function (Eq. 10), $S\big|_R(D)$ for $D = 10$ is

$$S\big|_R(10) = \exp(-H(10)) = \exp\Big(-\sum_{i=1}^{10} h(i)\Big) = \exp\Big(-\big(\frac{27}{25} + 5 \times \frac{1}{8} + 4 \times \frac{1}{24}\big)\Big) \doteq 0.154.$$

We now have the restricted conditional survival function $\frac{S\big|_R(t)}{0.154}$. Solving the restricted function for $t = 12$ results in

$$S\big|_R(12) = \exp(-H(12)) = \exp\Big(-\sum_{i=1}^{12} h(i)\Big) = \exp\Big(-\big(\frac{27}{25} + 5 \times \frac{1}{8} + 6 \times \frac{1}{24}\big)\Big) \doteq 0.142,$$

which finally gives $\frac{S\big|_R(12)}{S\big|_R(10)} \doteq 0.92$.

Application in fraud prevention – a high-probability (92 %) of login in two days from the same device's IP address after a ten-day use of the same IP address, may enforce a two-factor authentication if the actual login address is different.

IX. CONCLUSION

We presented an approach to predict the IP address assignment duration. The novelty is in the application of survival analysis, which is commonly used in healthcare. This concept allows censorship, which is useful when the datasets of prior address assignments do not cover a long period. Based on the created dataset of worldwide assignments over 6 years, we specified a general model of address lifetime expectancy. We further discussed the custom modelling for the cases when address survival prediction is made locally, i.e. when the hosts come from specific ISPs and ASs. As an example of use, we described a use-case in security, specifically in fraud prevention, where the prediction model is used as a trigger for two-factor authentication. The created dataset of addresses assignments is made publicly available at [1].

ACKNOWLEDGMENT

The survival analysis was based on the Python library Life-Lines [21].

REFERENCES

[1] D. Komosny and S. Rehman. (2020). *IP Address Assignment Durations*. Accessed: Jun. 27, 2020. [Online]. Available: http://dx.doi.org/10.21227/w2ez-w745
[2] R. Padmanabhan, A. Dhamdhere, E. Aben, K. Claffy, and N. Spring, "Reasons dynamic addresses change," in *Proc. ACM Internet Meas. Conf. (IMC)*, 2016, pp. 183–198.
[3] V. Bajpai, S. J. Eravuchira, and J. Schwnwalder, "Lessons learned from using the RIPE Atlas platform for measurement research," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 45, no. 3, pp. 35–42, 2015.
[4] G. C. M. Moura, C. Ganan, Q. Lone, P. Poursaied, H. Asghari, and M. van Eeten, "How dynamic is the ISPs address space? Towards Internet-wide DHCP churn estimation," in *Proc. IFIP Netw. Conf. (IFIP Netw.)*, May 2015, pp. 1–9.
[5] V. Mishra, P. Laperdrix, A. Vastel, W. Rudametkin, R. Rouvoy, and M. Lopatka, "Don't count me out: On the relevance of IP address in the tracking ecosystem," in *Proc. Web Conf.*, Apr. 2020, pp. 808–819.
[6] C. Wilcox, C. Papadopoulos, and J. Heidemann, "Correlating spam activity with IP address characteristics," in *Proc. INFOCOM IEEE Conf. Comput. Commun. Workshops*, Mar. 2010, pp. 1–6.
[7] G. Stringhini, T. Holz, S. Stone-Gross, C. Kruegel, and G. Vigna, "BOTMAGNIFIER: Locating spambots on the Internet," in *Proc. 20th USENIX Conf. Secur.* Berkeley, CA, USA: USENIX, 2011, pp. 1–32.
[8] H. J. Almohri, L. T. Watson, and D. Evans, "Predictability of IP address allocations for cloud computing platforms," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 500–511, 2020.
[9] S. Patil and N. Borisov, "What can you learn from an IP?" in *Proc. Appl. Netw. Res. Workshop*, Jul. 2019, pp. 45–51.
[10] D. Bojcic. (2017). *How Often do IP Addresses Change?* Vici Media Inc. Accessed: Jun. 27, 2020. [Online]. Available: https://www.vicimediainc.com/often-ip-addresses-change
[11] El Toro LLC. (2015). *How Long Does an IP Address Stay Attached to a Home or Business?* Accessed: Jun. 27, 2020. [Online]. Available: https://www.eltoro.com/how-long-does-an-ip-address-stay-attached-to-a-home-or-business
[12] T. Narten, R. Draves, and S. Krishnan, *Privacy Extensions for Stateless Address Autoconfiguration in IPv6*, document RFC4941, IETF, 2007.
[13] N. Vratonjic, K. Huguenin, V. Bindschaedler, and J.-P. Hubaux, "A location-privacy threat stemming from the use of shared public IP addresses," *IEEE Trans. Mobile Comput.*, vol. 13, no. 11, pp. 2445–2457, Nov. 2014.
[14] L. Vu, D. Turaga, and S. Parthasarathy, "Impact of DHCP churn on network characterization," *Perform. Eval. Rev.*, vol. 42, no. 1, pp. 587–588, 2014.
[15] M. Khadilkar, N. Feamster, M. Sanders, and R. Clark, "Usage-based DHCP lease time optimization," in *Proc. 7th ACM SIGCOMM Conf. Internet Meas. (IMC)*, 2007, pp. 71–76.
[16] O. Gasser, Q. Scheitle, P. Foremski, Q. Lone, M. Korczynski, S. D. Strowes, L. Hendriks, and G. Carle, "Clusters in the expanse: Understanding and unbiasing IPv6 hitlists," in *Proc. Internet Meas. Conf. (IMC)*, 2018, pp. 364–378.
[17] X. Fan and J. Heidemann, "Selecting representative IP addresses for Internet topology studies," in *Proc. 10th Annu. Conf. Internet Meas. (IMC)*, 2010, pp. 411–423.

[18] RIPE NCC. (2020). *RIPE Atlas Archive*. Accessed: Jun. 27, 2020. [Online]. Available: https://ftp.ripe.net/ripe/atlas/probes/archive/
[19] RIPE NCC. (2020). *RIPE Atlas Software Probe*. Accessed: Jun. 27, 2020. [Online]. Available: https://github.com/RIPE-NCC/ripe-atlas-software-probe
[20] J. Kishore, M. Goel, and P. Khanna, "Understanding survival analysis: Kaplan-Meier estimate," *Int. J. Ayurveda Res.*, vol. 1, no. 4, pp. 274–278, 2010.
[21] C. Davidson-Pilon, "Lifelines: Survival analysis in Python," *J. Open Source Softw.*, vol. 4, no. 40, p. 1317, Aug. 2019.

DAN KOMOSNY received the Ph.D. degree in teleinformatics, in 2003. He is currently a Professor with the Brno University of Technology, Czech Republic. His research interests include cybersecurity and digital forensics. He lecturers courses dealing with operating systems (UNIX/Linux) and IP networks (Cisco Networking Academy).

SAEED UR REHMAN received the Ph.D. degree in electrical and electronic engineering from The University of Auckland, New Zealand, in 2015. He is currently a Senior Lecturer of cybersecurity with Flinders University, Australia. His research interests include physical layer security and privacy-aware embedded systems.