Nonparametric Statistical Methods to Analyze the Internet Connectivity Reliability

Dimitri Papadimitriou
Alcatel-Lucent Bell Labs
Antwerp, Belgium
Email: dimitri.papadimitriou@alcatel-lucent.com

Davide Careglio
Technical University of Barcelona
Barcelona, Spain
Email: davide.careglio@ac.upc.edu

***Abstract*—Facing computational complexity when modeling network reliability by means of parametric models and corresponding statistical methods, in the present study, we apply nonparametric statistical methods, such as the Kaplan-Meier survival probability estimator and the mean cumulative function, to characterize the dynamic properties (in particular, the stability properties) of the Internet routing paths and their relationship with the corresponding forwarding path(s). Providing systematic methodology for quantifying these properties aims at enabling reliability assessment of the Internet connectivity (also referred to as reachability in computer networking). The motivation for studying the dynamic properties (in particular, the stability properties) of the Internet routing paths and their relationship to forwarding paths stems from three main reasons. The first translates the fact that transient but frequent changes in the spatio-temporal properties of routing paths may affect the performance and operating conditions of the corresponding forwarding paths; hence, their reliability. The second reason is that frequent instabilities when observed for the same (subset of) path(s) that can be attributed to a spatially localized portion(s) of the Internet may reveal that the underlying physical topology is more prone to failures; hence, showing limited reliability. The third results from the increasing operational need to provide for a longer term estimation of the Internet routing-forwarding system performance and operating conditions using well-proven statistical analysis accounting for recurrence of events and correlation between instability events.**

I. Introduction

The Internet as most prominent example of complex (engineered) system or network leads one to consider data- instead of model-driven statistical methods. The importance of nonparametric methods in statistics has grown significantly since their inception in the mid-1930s. Requiring few or no assumptions about the populations from which data are obtained, they have emerged as the preferred methodology among statisticians and researchers performing data analysis. Today, nonparametric statistical techniques are being applied to an ever-widening variety of experimental designs in the social, behavioral, biological, and physical sciences. The need is evident for statistical procedures that enable to process data of relatively low quality, from small samples, on variables about which little or even sometimes nothing is known concerning their distribution. Specifically, nonparametric methods were developed to be used in cases when nothing is known about the parameters of the variable of interest in the population. Unlike parametric statistics, nonparametric statistics make no assumptions about the probability distributions of the variables under study. Indeed, these methods which include both descriptive and inferential statistics do not rely on the estimation of parameters such as the mean or the variance, describing the distribution of the variable of interest in the population. Consequently, these statistical methods are sometimes called parameter-free methods or distribution-free methods.

For what concerns topology failures, the generalization of the Weibull distribution is often considered when modeling the reliability of engineered systems and their components (e.g., physical links) by their failure probability at specific time as well as their failure rate variation over time. Nevertheless its application to large-scale networks where each link failure rate depends on different parameters values leads to consider multivariate joint distributions that is not simply the product of the individual distributions. This method finds applicability when modeling simultaneous and/or correlated failures together with their probability distribution, including node failures and common resource failures which lead to the failure of numerous network paths. In this paper, we apply instead nonparametric statistical methods, namely the Kaplan-Meier survival probability estimator and the mean cumulative function, to characterize the dynamic properties (in particular, the stability properties) of the Internet routing paths and their relationship with their corresponding forwarding path(s). By stability we refer in the present paper to the spatial changes affecting the sequence of (abstract) nodes and edges of the path from the same source to the same destination. The motivation for studying the dynamic properties of the Internet routing paths and their relationship(s) to forwarding paths stems from three main reasons. The first translates the fact that transient but frequent changes in the spatio-temporal properties of routing paths may affect the performance and operating conditions of the corresponding forwarding paths; hence, their reliability. The second reason is that frequent instabilities when observed for the same (subset of) path(s) that can be attributed to a spatially localized portion(s) of the Internet may reveal that the underlying physical topology is more prone to failures; hence, showing limited reliability. The third results from the increasing operational need to provide for a longer term estimation of the Internet routing-forwarding system performance and operating conditions using well-proven statistical analysis accounting for recurrence of events and correlation between various events. In this context, providing a nonparametric method to analyze these properties aims at providing a quantitative evaluation of the reliability of the Internet connectivity (also referred to as reachability in

 DOI 10.1109/CQR.2015.7129083

computer networking) and its evolution over time (prediction) without requiring to infer the parameters of the distribution characterizing network failure probability and rate.

On the other hand, reliability and resilience are intimately related: resilient networks are designed such as to provide the capability to improve their fault tolerance and, as a result, their reliability. A resilient system is one that can withstand a number of (sub-)system and components faults or failures while continuing to provide and maintain an acceptable level of performance in the face of various faults and challenges to normal operation. Resilience may indeed be defined as systems ability to either absorb or tolerate change without losing one's peculiar traits or expected behaviors by considering that systems exist close to a stable steady-state. Consequently, resilience can thus be seen as the ability of a system to return to the steady state following a perturbation, and, in the context of this paper, it refers to the characteristic of being able to adapt, absorb or tolerate change, perturbation or faults without losing functionality and its expected behavior.

The remainder of the paper is organized as follows. In Section II, we introduce the base definitions and concepts used throughout this paper. Section III outlines the prior work and related methods considered in reliability analysis together with its application to communication networks; we also detail in this section the main objectives and contribution of this paper. Section IV documents the nonparametric statistical methods considered in this paper, namely the Kaplan-Meier survival probability estimator and the mean cumulative function. The results obtained by the application of the nonparametric statistical methods are documented in Section V together with their analysis. Finally, Section VI draws conclusion from this study and outlines possible future work.

II. Preliminaries

Consider a network topology modeled by an undirected graph $G = (V, E)$ where, the vertex set $V(|V| = n)$ represents the finite set of nodes and the edge set $E(|E| = m)$ represents the finite set of links. For $u, v \in V$, the loop-free path $p(u, v) \in \mathcal{P}$ from vertex u to v is defined as the finite sequence $[v_0(= u), v_1, \ldots, v_{i-1}, v_i, \ldots, v_p(= v)]$ such that the vertex v_{i-1} is adjacent to v_i, $\forall (v_{i-1}, v_i)_{i=1,\ldots,p} \in E$. Following this definition a path instability (or perturbation) is characterized by a change (or interruption if detectable) in the sequence of vertices along the path $p(u, v)$ from the same source u to the same destination t. We refer the reader to [5], [8] and [10] for additional details concerning path stability and related metrics as well as computational procedures. Further distinction is made between topological, forwarding and routing paths. A topological path $p(u, v)$ from vertex u to v denotes a path determined out of the topology graph G. The distance between vertex u and v is defined by the shortest distance (topological) path defined on the network graph G. The routing path $r(u, t)$ denotes the (not necessarily shortest) path towards destination $t \in \mathcal{D}$ (the set of destinations) as produced at node u by the distributed routing algorithm using as input the information of the graph G. The routing topology defines thus a sub-graph H of G representing the actual nodes and links along the paths as selected by the routing algorithm (and stored in local routing tables). A forwarding path $f(u, t)$ denotes the path followed by the data traffic from node u to destination t. The forwarding path is derived at each node from the local routing table information. The forwarding topology defines thus a sub-graph H' of G representing the actual nodes and links as selected by each routers forwarding decision. Usually, the sub-graphs H and H' are not required to be identical (a routing table entry may exist without a corresponding forwarding entry).

Let T be a continuous random variable representing the failure time or lifetime of a physical system with cumulative distribution function $F(t)$ at time t. The probability that the system will fail by time t is given by the following equation

$$F(t) = P[T \leq t] = \int_0^t f(x)\, dx \tag{1}$$

In this equation, the failure probability density function (p.d.f.) $f(t)$ is defined as the expected number of failures experienced in a given time interval

$$f(t) = \frac{dF(t)}{d(t)} \tag{2}$$

The probability that the system survives at least until time t is given by the reliability function $R(t)$ (or survival function $S(t)$). This function, which captures the probability that the system will survive beyond a specified time t, relates to $F(t)$ by the following equation

$$R(t) = P[T > t] = 1 - F(t) = \int_t^\infty f(x)\, dx \tag{3}$$

$$R'(t) = \frac{d}{dt} R(t) = -f(t) \tag{4}$$

Assume the system survived up to time t (the failure event has not occurred before time t) and fails by $t + \Delta t$ (the failure occurs in the interval $[t, t + \Delta t]$), the average failure rate $\lambda(t)$ in the time interval Δt is given by

$$\lambda(t) = \frac{F(t + \Delta t) - F(t)}{\Delta t R(t)} = \frac{R(t) - R(t + \Delta t)}{\Delta t R(t)} \tag{5}$$

The hazard function $h(t)$ which represents the instantaneous failure rate at time t, given that the system survived until time t, is defined as

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t R(t)} = \lim_{\Delta t \to 0} \frac{R(t) - R(t + \Delta t)}{\Delta t R(t)} \\ &= \frac{f(t)}{R(t)} = \frac{-R'(t)}{R(t)} = -\frac{d(\ln R(t))}{dt} \end{aligned} \tag{6}$$

We will refer to it to as the failure rate function $h(t)$ to conform to the common usage in many references (e.g., [11]). Following this definition, the cumulative failure rate function $H(t)$ is defined as $\int_0^t h(x)\, dx$.

III. Prior Work, Related Methods and Contribution

The Weibull distribution (and its variants such as the Weibull-logarithmic distribution [13]) is commonly used in the context of survival analysis and reliability engineering to model and estimate the characteristics and behavior of equipment and systems such as their failure probability at specific time, and their failure rate variation over time. Consider for instance the 2-parameter Weibull probability density function

with scale parameter $b > 0$ (or slope), and shape parameter $c > 0$ (or characteristic life):

$$f(t) = \frac{dF(t)}{d(t)} = \frac{c}{b}(\frac{t}{b})^{c-1} \exp{(-\frac{t}{b})^c} \quad (7)$$

The equation for the 2-parameter Weibull cumulative density function $F(t)$ is given by:

$$F(t) = P[T \leq t] = \int_0^t f(x)\, dx = 1 - \exp{(-\frac{t}{b})^c} \quad (8)$$

The probability of the system surviving until time t is given by the function $R(t)$ referred to as the survival function or reliability function:

$$R(t) = P[T > t] = 1 - F(t) = \int_t^\infty f(x)\, dx = \exp{(-\frac{t}{b})^c} \quad (9)$$

The Weibull instantaneous failure rate $h(t)$ is given by:

$$h(t) = \frac{f(t)}{R(t)} = \frac{c}{b}(\frac{t}{b})^{c-1} \quad (10)$$

Despite its popularity, and wide applications, the traditional 2- or 3-parameter Weibull distribution is unable to capture the behavior of a lifetime data set that has a non-monotonic failure rate function. For this reason, many distributions were proposed to overcome this deficiency. Following the fundamental relationship between the reliability function $R(t)$ and its corresponding cumulative failure rate function $H(t)$, $R(t) = exp(-aH(t))$ with $a > 0$, various generalized Weibull models have been proposed in the reliability literature [12]. With a suitable choice of $H(t)$, and its parameter, we can obtain a bathtub shaped failure rate distribution. Generalized Weibull distributions provide a reference of choice when modeling the reliability of engineered systems and their components (e.g., physical links) by their instantaneous failure probability as well as their failure rate variation over time. Nevertheless their application to large-scale networks where each link failure rate depends on different shape and scale parameters values leads to consider multivariate joint distributions that are not simply the product of the individual distributions when modeling simultaneous and/or correlated failures, e.g., node failures, or common resource failures. Indeed additional parameters are required to account for coupling effects $\nu > 0$ but also different time thresholds $\tau \geq 0$ (because all links are rarely installed at the same time) in order to model the joint survival distribution of sets of network components. More precisely, the joint survival distribution $R_{\mathcal{K}}(t)$ of a set $\mathcal{K}$ comprising k components (e.g., k network links) with individual failure rates λ_k is defined by the following expression (see [14]):

$$R_{\mathcal{K}}(t) = \exp\left(\tau_k^\nu - \left[\tau_k + \sum_{i=1}^{k}(\lambda_i t_i^{c_i})\right]^\nu\right) \quad (11)$$

Assuming that the distribution parameters can be inferred out of the observations obtained via network monitoring (which is often not the case in practice), the resulting model is thus computationally intractable without simplifying assumptions. In order to characterize the reliability of the Internet connectivity it becomes thus attractive to apply nonparametric statistical methods to the data obtained from the observation of the forwarding and the routing paths behavior as well as their relationships. Instead of modeling and analyzing their reliability starting from the joint failure probability of the different components underlying the network topology (thus, following a bottom-up structural approach), this paper approaches the crucial question of the reliability properties of the Internet connectivity and subsequent root cause analysis from the top-down statistical perspective.

IV. Non-parametric Statistical Analysis and Methods

A. Kaplan-Meier Survival Probability Estimation

The Kaplan-Meier estimator provides non-parametric maximum likelihood estimate $\hat{S}(t)$ of the survival function $S(t)$. Assume k distinct event times $t_1 < \ldots < t_i < \ldots < t_k$. Let $d_1 < \ldots < d_i < \ldots < d_k$ denote the number of failures (or deaths) that occurred at event time t_i. For each event time t_i, the risk set n_i represents the number of entities at risk just before time t_i which consists of the original sample minus all those who have already been censored or experienced the event before time t_i. It is only those surviving cases that are still being observed (have not yet failed or been censored) that are "at risk" of an (observed) failure. When there is no censoring, n_i represents the number of survivors prior to time t_i. With censoring, n_i is the number of survivors minus the number of losses (censored cases). Note that censoring means that the total survival time for that subject cannot be accurately determined. This can happen when something negative for the study occurs, such as the subject drops out, is lost to follow-up, or required data is not available or, conversely, something good happens, such as the study ends before the subject had the event of interest occur, i.e., they survived at least until the end of the study, but there is no knowledge of what happened thereafter. Thus, censoring can occur within the study or terminally at the end.

Intuitively, the probability of surviving beyond time t_{i+1}, $S(t_{i+1}) = P(T > t_{i+1})$ depends conditionally on the probability of surviving beyond time t_i, i.e., $S(t_i) = P(T > t_i)$. By using this recursive concept, one can iteratively build a numerical estimate $\hat{S}(t)$ of the true survival function $S(t)$. Formally, it is obtained by applying the conditional probability formula $P(A \cap B) = P(A) \times P(B|A)$, where A denotes the event to survive to time t_i, B survive from time t_i up to some time t before t_{i+1}, and the event $A \cap B$ to survive to beyond time t before t_{i+1}. The estimated probability $P(T > t)$ to survive to time $t \in [t_i, t_{i+1})$ is given by the Kaplan-Meier estimator $\hat{S}(t)$, a right-continuous step function defined as:

$$\hat{S}(t) = (1 - \frac{d_1}{n_1})(1 - \frac{d_2}{n_2}) \cdots (1 - \frac{d_i}{n_i}) \quad (12)$$

$$= \prod_{j=1}^{i}[1 - \frac{d_j}{n_j}] = \prod_{j:t_j \leq t}[1 - \frac{d_j}{n_j}] \quad (13)$$

$$= \prod_{j=1}^{i}[\frac{n_j - d_j}{n_j}] = \prod_{j:t_j \leq t}[\frac{n_j - d_j}{n_j}] \quad (14)$$

The ratio $\frac{d_j}{n_j}$ in (13) represents the proportion that failed at the event time t_j and $(1 - \frac{d_j}{n_j})$ the proportion that survived to

the event time t_j. In (14), the numerator defines the number of surviving entities after event time t_j and the denominator n_j the number of surviving entities at risk (a.k.a. risk set) in the interval just prior to event time t_j. Following this definition, the Kaplan-Meier estimator $\hat{S}(t)$ can be regarded as a point estimate of the survival function $S(t)$ at any time t.

The Kaplan-Meier survival probability curve is defined as the probability of surviving in a given length of time while considering time in many small intervals. The Kaplan-Meier procedure is not limited to the measurement of survival in the narrow sense of dying or not dying. It can also be used to estimate the time-defined probabilities for the failure of a device of a certain type; or alternatively, to estimate the time-defined probabilities for some particular type of success (e.g., being repaired after experiencing failure). An important advantage of the Kaplan-Meier curve is that the method can take into account some types of censored data, particularly right-censoring, which occurs if an entity is withdrawn from a study, i.e. is lost from the sample before the final outcome is observed. On the plot of the Kaplan-Meier curve, small vertical tick-marks indicate losses, where a entity's survival time has been right-censored. Note also that when no truncation or censoring occurs (which is the case for repairable systems), the Kaplan-Meier curve is the complement of the empirical distribution function.

B. Recurrent Event Data Analysis (RDA)

Recurrent event data arise when observed entities possibly experience more than one event during the observation periods. A key characteristic is that some observations may be correlated, a failure event is observed through the occurrence of multiple link, node or path failure events (on which they rely). Moreover, the temporal trajectories of observed data are often very complex to determine and their statistical modeling leads to generalized multivariate distributions difficult to use in practice for analytic or predictive purposes. Consequently, parametric statistical models may not be flexible enough to capture their main features and (stochastic) temporal networks where the sequence of activation times is a stochastic model that preserves the observed inter-event distribution difficult to apply in practice; instead, nonparametric (or semi-parametric) statistical models are particularly attractive in such study. In these models, the mean structures are modeled non-parametrically (or semi-parametrically) and the distributional assumptions are assumed to be non-parametric.

Recurrent Event Data Analysis (RDA) is commonly used in various engineering fields and is particularly useful when performing reliability analysis of repairable systems. Conventional Life Data Analysis (LDA) assumes that events (failures) are independent and identically distributed (i.i.d) whereas in certain situations, the events are dependent and not identically distributed (common property of repairable system data). Moreover, we are interested in modeling the number of occurrences of events over time rather than the length of time prior to the first event occurrence as in LDA; indeed, the latter typically focuses on time to event occurrence data. Non-parametric RDA provides a non-parametric graphical estimate of the mean cumulative number of recurrences of events versus time. This non-parametric analysis method relies on the Mean Cumulative Function (MCF).

When analyzing repairable systems, the simplest plot that can be constructed is a cumulative plot (staircase function), which graphs the number of recurrences of events over time. It is also possible to represent the behavior of the set of events by computing the average of the cumulative number of recurrences of events over time. This average is referred to as the Mean Cumulative Function (MCF). Compared to the Mean Time Between Failure (MTBF), the MCF method provides the following advantages i) the MCF is more adequate to represent the event rate because it makes no distributional assumptions (nonparametric method) and ii) the MCF is also more informative because it provides trends in the event rate as a function of time. The MCF reveals also one important aspect in reliability analysis by providing an estimation of the expected rate of events, i.e., the (cumulative) number of perturbation events that forwarding and routing paths experienced over time.

To compute the MCF $M(t_i)$ (more precisely, the estimation of the MCF) one proceeds as follows. At each observation time t_i, the number of events n_i that occurred since the previous observation time t_{i-1} is recorded. These events are recurrent (i.e., non-fatal like failure events followed by restoration) and are assumed to occur randomly. The number of events n_i is divided by the number ρ_{i-1} of pairs that are observable at time t_{i-1} (with ρ_1 set to the total number of initially observable number of entities). One then computes the MCF estimate using the formula:

$$M(t_i) = \frac{n_i}{\rho_{i-1}} + M(t_{i-1}) \tag{15}$$

$$M(t_1) = \frac{n_1}{\rho_0} \tag{16}$$

It is important to note that all observation intervals are taken to be nonrandom, identical for all observations and of equal length. Moreover, since the MCF is an estimate, confidence bounds can also be computed. The shape of the MCF plot can reveal several important properties about the behavior of the recurrent events under consideration in a reliability study. The MCF vs. time (age) curve can be numerically differentiated to obtain the slope, called the recurrence rate. From the shape of the MCF plot, one can then derive the following interpretation assuming that instability events induce transient changes in paths properties that affect their performance and operating conditions (hence, their reliability):

- Constant recurrence rate: the MCF plot increases monotonically, the slope of the MCF remains constant; thus events under consideration occur at constant rate.
- Increasing recurrence rate: the MCF plot is convex (the slope of the MCF increases), the recurrence rate increases over time which reveals system performance degradation over time.
- Decreasing recurrence rate: the MCF plot is concave (the slope of the MCF decreases), the recurrence rate decreases over time which reveals maintenance improvement over time (decreasing repair rate).

V. APPLICATION

Following the seminal paper of V.Paxson [1], prominent research efforts [2], [3], [4], [5] have been conducted to

understand the dynamic properties of the Internet routing paths and, in particular, the root cause(s) of their instabilities. In reactive routing, active measurements of network-layer path characteristics such as reachability, loss, and latency are used in combination with passive traffic monitoring to decide which paths are better; the differences are in how they take advantage of alternate paths. The proposal to improve the availability of the Internet connectivity by means of reactive routing instead of adaptive routing (which relies on routing information changes to select between paths) led to study the relationships between forwarding paths and routing paths properties. As observed by Z.M.Mao et al. [6], routing and forwarding paths do not always match due in particular to route aggregation and forwarding anomalies. In [7], N.Feamster et al. have reported several results on the relationships between forwarding path failure and Internet routing system instability induced by messaging. This study further observed that i) forwarding path failures often precede BGP instability as BGP messages may indeed follow the appearance of a failure due to the slow adaptation of the BGP routing system upon failure occurrence but ii) BGP instability may also precede forwarding path failures when routing path changes signal (BGP messages) external events such as network maintenance or where routing protocol engines themselves are the cause of such failures.

The present study aims by applying nonparametric statistical methods to characterize and to analyze the dynamic properties of the forwarding and the routing paths as well as their relationships. By the latter we mean more precisely to determine whether routing paths follow mainly the perturbations experienced by the forwarding paths or vice versa. In particular, we show that simple causality effect as assumed in [7] is not verified anymore. Our analysis determines that the main cause of perturbation results from the forwarding plane, confirming the results reported in [7] and corroborating the assumption that the dynamic properties of the routing system are mainly driven by its adaptation to the forwarding system.

The present study relies on the detection and identification of perturbation events following the methods and procedures documented in [8] based on the stability criteria and metrics introduced in [5], [9], [10]. The routing path information is extracted from BGP datasets provided by the RouteViews project. These datasets collected over 50 days from the monitored BGP routers, comprise the following information: i) the complete Routing Information Base (RIB) entries (updated every two hours) and ii) the received BGP routing updates received from peering AS (separated in files recorded every 15 minutes). The forwarding paths information is extracted from the data recorded by the RADAR tool. The measurements carried out by this tool are traceroute-like probes initiated from a set of monitoring nodes. Such probes target a large set of IP address prefixes distributed across the Internet. Based on these measurements, the RADAR tool builds ego-centered views of the forwarding topology, i.e., the initiating router collects traces along the forwarding paths that it probes. A subset of the forwarding paths traced by the RADAR tool corresponds to the routes obtained from the RouteViews dataset. Consequently, a subset of the monitored routing paths is also monitored by the RADAR tool. In total, the analyzed dataset includes a bit less than 1000 forwarding path - routing path pairs covering a monitoring period of 50 days.

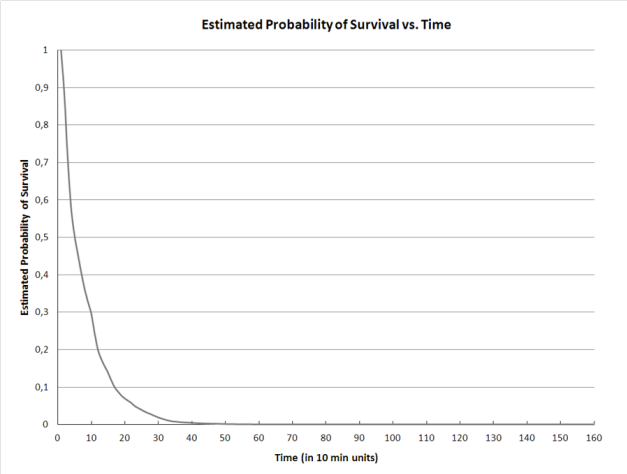

Fig. 1. Kaplan-Meier survival probability curve (Forwarding paths)

A. Kaplan-Meier Survival Probability Estimation

Applying the procedure without censoring as documented in Section IV-A to our dataset yields the Kaplan-Meier curve of Fig. 1. This figure plots the estimated probability of survival (meaning in the present case, the number of paths surviving beyond time t_i perturbation events occurring at that time) which drops quickly. In other terms, the perturbations affecting directly (or indirectly) forwarding paths are common events leading rapidly to (within 10 hour) to a complete connectivity unavailability if no reactive (or proactive) action are being taken to restore connectivity. Note that the absence of censoring can be seen from the absence of tick mark which would represent the time one of the elements in the dataset would have been censored. On the other hand, the probability of survival for the routing paths follows the same type of curve than the one observed for the forwarding paths but with a time shift. The main observation that can be drawn from Fig. 2 is that after the first 10 hours of observation, the survival probability slowly decreases to reach about 90% whereas the corresponding estimation for the forwarding paths already dropped to less than 1%. During that period, it is also interesting to observe that the latter curve which is exponentially decreasing seems to indicate a second order effect on the survival probability estimation for the routing paths. From this curve, we can also observe that major routing perturbation events at 100 and 700 (along the X-axis) seriously affect the survival probability estimation of the routing paths whereas between these major events this probability decreases rather slowly (< 0.001) compared to the behavior observed for the forwarding paths.

B. Recurrent Data Analysis (RDA) and Mean Cumulative Function (MCF)

Applying the procedure documented in Section IV-B to our dataset yields the MCF plot of Fig. 3. This result indicates that both forwarding and routing paths experience instabilities at constant recurrence rate. However, the rate experienced by the forwarding paths is about 10 times higher than the rate experienced by the routing paths. In other terms, and following these results, the main source of perturbation affecting the reliability of the Internet connectivity would be caused the

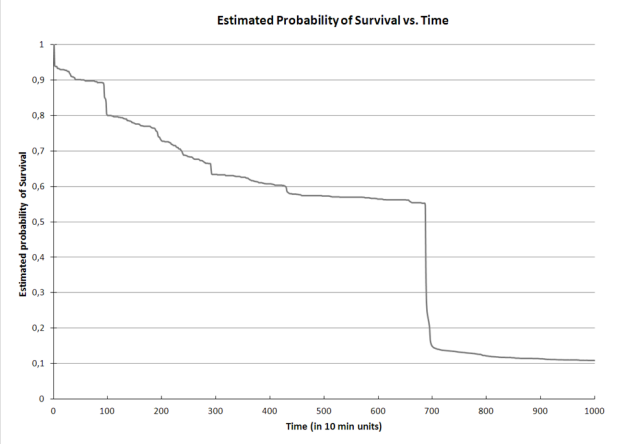

Fig. 2. Kaplan-Meier survival probability curve (Routing paths)

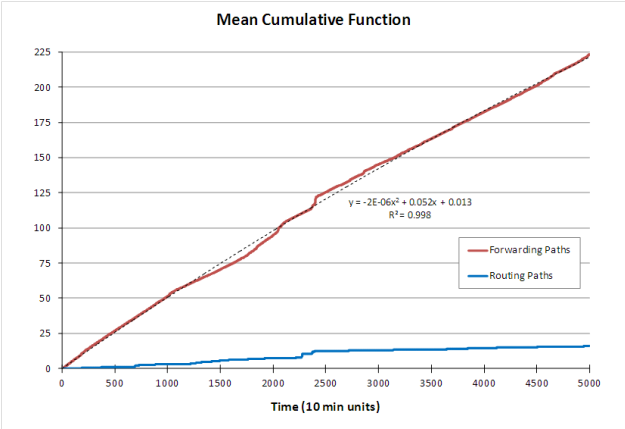

Fig. 3. Mean Cumulative Function (MCF)

forwarding plane. It is important to remember here that the data being analyzed cover 1000 forwarding-routing path pairs; assuming that the number of active paths in the Internet is about 512k as of July 2014, we get a representative sample size of the total population to obtain a confidence level of 95% and a confidence interval of ± 3. We would need approximately 16k pairs to reach a confidence level of 99% with a confidence interval of ± 1. The main limit to generalization comes thus mainly from the nonrandom selection of the location where the datasets have been obtained.

VI. Conclusion

By applying nonparametric statistical methods, namely the Kaplan-Meier survival probability estimator and the mean cumulative function, the present study aims at exemplifying their applicability for the characterization of the dynamic properties (in particular, the stability properties) of the Internet routing paths and their relationship with the corresponding forwarding path(s). These properties indeed directly influence the reliability of the Internet connectivity (also referred to as reachability in computer networking). In particular, we show that by using these methods, simple causality effects between forwarding and routing paths unavailability as identified about 10 years ago in [7] are not verified anymore. Our analysis determines that the main source of connectivity perturbations is caused by the forwarding plane instabilities (which experiences a 10x higher recurrence rate compared to the routing plane). This observation confirms the results reported in [7] and corroborates the assumption that the dynamic properties of the routing system are mainly driven by its adaptation to the forwarding system. As reported in [8], forwarding path instabilities induce routing path instabilities while the corresponding forwarding path remains unstable. This result suggests thus that reactive-like routing systems are now in place but 50% of their decisions tend to delay the convergence of forwarding paths (instead of only delaying convergence of routing paths through adaptive routing as reported by Z.M.Mao et al. [6]). It shows also that the causality effect assumed in [7] does not find any more a simple explanation as forwarding paths become the dominant source of instability affecting the reliability of the Internet connectivity. Obviously, reproducing the same nonparametric statistical procedures on similar datasets obtained from randomly selected locations would enable to further generalize the outcomes of this study. As part of our future work, we will also extend the analysis method to localize and characterize intra-AS instabilities.

References

[1] V. Paxson, *End-to-End Routing Behavior in the Internet*, IEEE/ACM Transactions on Networking, vol.5, no.5, pp.601-615, 1997.

[2] C. Labovitz, R. Malan, and F. Jahanian, *Origins of Internet Routing Instability*, Proc. of IEEE INFOCOM 1999, New York (NJ), March 1999.

[3] C. Labovitz, A. Ahuja, A. Bose, and F. Jahanian, *Delayed Internet Routing Convergence*, IEEE/ACM Transactions on Networking, vol.9, no.3, pp.293–306, June 2001.

[4] T. Griffin, F. B. Shepherd, and G. Wilfong, *The Stable Paths Problem and Interdomain Routing*, IEEE/ACM Transactions on Networking, vol.10, no.1, pp.232–243, April 2002.

[5] D. Papadimitriou, A. Cabellos, and F. Coras, *Path-vector Routing Stability Analysis*, Proc. of 13th Workshop on MAthematical Performance Modeling and Analysis (MAMA), ACM SIGMETRICS 2011, San Jose (CA), June 2011.

[6] Z. M. Mao, R. Govindan, G. Varghese, and R. Katz, *Route Flap Damping Exacerbates Internet Routing Convergence*, Proc. of ACM SIGCOMM 2002, Pittsburgh (PA), August 2002.

[7] N. Feamster, D. Andersen, H. Balakrishnan, and F. Kaashoek, *Measuring the Effects of Internet Path Faults on Reactive Routing*, Proc. of ACM SIGMETRICS 2003, San Diego (CA), June 2003.

[8] D. Papadimitriou, D. Careglio, F. Tarissan, and P. Demeester, *Method of reliability and availability analysis From the dynamic properties of routing and forwarding paths*, Proc. of 5th IFIP/IEEE Int'l Workshop on Reliable Networks Design and Modeling (RNDM),September 2013.

[9] D. Papadimitriou, D. Careglio, F. Tarissan, and P. Demeester, *Internet routing paths stability model and relation to forwarding paths*, Proc. of 9th Int'l Conference on the Design of Reliable Communication Networks (DRCN), Budapest, Hungary, March 2013.

[10] D. Papadimitriou, A. Cabellos, and F. Coras, *Stability metrics and criteria for path-vector routing*, Proc. of IEEE International Conference on Computing, Networking and Communication (ICNC) 2013, San Diego (CA), January 2013.

[11] I. Bazovsky, *Reliability theory and practice*, Englewood Cliffs (NJ), Prentice-Hall, 1961.

[12] D. Murthy, M. Xie, and R. Jiang, *Weibull models*, Hoboken (NJ), John Wiley & Sons, Inc., 2004.

[13] R. Ciumara, V. Preda, *The Weibull-logarithmic distribution in lifetime analysis and its properties*, In L. Sakalauskas, C. Skiadas and E.K. Zavadskas (Eds.) Applied Stochastic Models and Data Analysis, The XIII International Conference, Selected papers. Vilnius, 2009.

[14] M. Crowder, *A Multivariate Distribution with Weibull Connections*, Journal of the Royal Statistical Society, Series B (Methodological), Vol.51, No.1, pp.93-107, 1989.