Removing bias due to finite measurement of dynamic networks

Lamia Benamara[a], Clémence Magnien[a]

[a] *UPMC Univ Paris 06, UMR 7606, LIP6, F-75252, Paris, France*
CNRS, UMR 7606, LIP6, F-75252, Paris, France
Email: firstname.surname@lip6.fr

Abstract

Characterizing accurately the dynamics of evolving networks (such as P2P systems, the internet topology, ...) is a difficult task, because several factors can introduce a bias in the observed properties. In particular, the fact that we can observe a given system only for a finite duration introduces a bias, because events occurring before or after the observation are missed. Although this bias tends to decrease when the observation window length increases, it is difficult to quantify its importance, or know how fast it decreases.

Here, we introduce a general methodology that allows to know if the observation window is long enough to characterize a given property.

We apply this methodology to the study of several properties in a large P2P system, using two different and complementary datasets. We show that an observation window that is too short does indeed induce a bias, and that our methodology allows to detect this. We also show that there is no overall satisfying duration for observing a given system. While some properties can be characterized with a given observation window length, others cannot be characterized at all in our datasets, either because the measurement was not long enough, or because the property is intrinsically not stationary. In either case, these properties cannot be trusted.

Keywords:
Dynamics, metrology, bias, measurement, P2P.

1. Introduction.

Many systems are naturally dynamic. For instance in the internet, routers, AS and/or links between them are created or deleted [9, 10]; in P2P networks, users join or leave the system [13, 12, 8], and exchange different files at different times; in online social networks users may create or delete accounts, or cease to be active, and create or delete connections with other users [15].

In all these cases, understanding the dynamics of the system is a key issue. However, accurately measuring this dynamics is a difficult task. In particular, the fact that the observation window is necessarily finite induces a bias in the observations [11, 13, 12]. Though this bias tends to decrease when the observation window length increases, it is difficult to quantify it in practice, and know whether it is negligible or not.

Another problem is that a small observation window may not be representative of the whole behavior of the system. For instance, measuring the activity in a P2P system during one hour is not enough to capture fully the dynamics of user usages, because of day/night activity variations for instance. However, it is not *a priori* clear whether one day, or two, or one week, is long enough.

In this paper, we introduce a new methodology that allows to rigorously characterize dynamic metrics in real-world dynamic systems. This methodology is different and complementary to other methodologies existing in the literature [13, 12], and has two main advantages:

- it allows to determine if the observation window length was sufficient for a rigorous characterization;
- it can be applied to any property characterizing the dynamics of a system.

To illustrate the relevance of this methodology, we apply it to the study of several properties in the *eDonkey* P2P system. We use two different datasets which provide complementary information.

This document is organized as follows. In Section 2, we introduce our methodology and present the datasets we use. In Section 3 to 7 we apply our methodology to the study of several properties describing the system. We present related work in Section 8, and our conclusions and future work in Section 9.

2. Methodology and Data

2.1. Methodology

Suppose we start observing a dynamic graph at a time t, for a duration l. We denote by $W_{t,l}$ this observation window. We are faced with two problems if we want to characterize the graph's dynamics from the observation of $W_{t,l}$. First, l must be long enough for $W_{t,l}$ to be *representative*. For instance, it seems hopeless to characterize rigorously the activity in a P2P system after observing it for a single hour: at the very least, this does not allow to observe the activity variation according to the time of the day. Second, even if it is representative, the fact that l is *finite* still induces a bias in the observations. Events occurring before t or after $t+l$ are not observed, which prevents from characterizing accurately some quantities (for instance, session lengths, or time correlations between different events). An important point to observe is that the longer the measurement period, the smaller the bias induced.

Our methodology addresses these two issues at the same time. Intuitively, it aims at deciding if the measurement period $W_{t,l}$ is long enough to characterize a given property P, i.e. if the bias induced by its finiteness on the observed property is negligible. If the window $W_{t,l}$ is long enough, then if we use a longer window of length $l + x$, the observed property does not change: $P(W_{t,l}) = P(W_{t,l+x})$.

In order to decide when a window is long enough, we use windows of increasing length W_{t,l_1},W_{t,l_2}, ..., W_{t,l_n} ($l_1 < l_2 < ... < l_n$). By studying how the observed property $P(W_{t,l_1})$,$P(W_{t,l_2})$, ...$P(W_{t,l_n})$ evolves as a function of l, we determine if it is correctly evaluated or not: if it fluctuates or varies greatly as l increases, then P is certainly not accurately evaluated. Indeed, a shorter or longer observation window would have yielded a different value. Instead, if P tends to become stable as the window length l increases, then it is probably accurate.

Finally, an important point is that characterizing a property P only makes sense if it is stationary, i.e. if P does not evolve while the measurement is under progress. Notice however that if it is not stationary, our methodology will not be able to provide a characterization: the observed property P will not become stable when the observation window length l increases. If it does become stable, this means both that $W_{t,l}$ is long enough, and that P is stationary [1].

Notice that, depending on the property studied, other types of bias can occur, see for instance [13]. In our context, some come from the identification of users and their sessions. We do our best to deal with them in a rigorous way, as we detail in the following sections. However, we stress on the point

[1] Note that the system may be stationary with respect to a given property P and not another one P'; in such a case our methodology will provide a characterization for P and not for P'.

that our goal here is not to address all kinds of biases at the same time, but to exhibit the role played by the observation window length.

Here, most of the properties we study are distributions. In general we will denote a distribution with a subscript k to indicate that it is a function of k, e.g. P_k. To study how an observed distribution P_k evolves with the length of the observation window, we will first plot the observed distributions $P_k(W_{t,l})$ for different values of l. We note that we take t as the beginning of our measurement period, therefore we set $t = 0$ in the following.

In order to confirm more formally the visual observations, we will also study two statistical indicators which quantify how close two distributions P_k and Q_k are to each other. The *Kolmogorov-Smirnov test*, or K-S test [3] compares two normalized cumulative (complementary or not) distributions P_k and Q_k. It is equal to the maximum, for all values k, of the distance between the two cumulative distributions: $\mathrm{KS}(P_k, Q_k) = \max_k |P_k - Q_k|$. It is always lower than 1, and the closer it is to 0, the more similar the two distributions are.

An important question raised by the K-S test is to know if the distributions differ by the resulting value at all points, or just at one point. In order to help us answering this question, we study the *Monge-Kantorovich distance*, or M-K distance [5] which is equal to the mean of the distance between the two (cumulative) distributions: $\mathrm{MK}(P_k, Q_k) = (\sum_k |P_k - Q_k|)/k_{\max}$. Two distributions that only differ in a single point will therefore have a high K-S test, but a small M-K distance. We use these indicators to study how the observed distribution $P(W_{0,l})$ evolves: we compute the K-S test (respectively the M-K distance) between $P(W_{0,l})$ and $P(W_{0,l_{max}})$, where l_{max} is the length of the longest observation window available for this dataset, and plot this as a function of l. Following [17], we also study the mean and standard deviation of $P_k(W_{0,l})$ as a function of l.

2.2. Data

We use two datasets: the first consists in the capture of the UDP traffic of a large *eDonkey* server [1]. It consists of the queries made by users (for lists of files matching certain keywords, or for providers for a given file), and of the server's answers to these queries. There are two types of queries. The first one are of the following form:

$$T \quad IP \quad L,$$

where T is the time at which this query was made, IP is the (anonymized) IP address of the user making this query, and L is a list of keywords describing the wanted file. The servers's answer is of the following form:

$$T \;\; IP \;\; (F_1, S_1) \;\; (F_2, S_2) \;\; ... \;\; (F_n, S_n),$$

where IP is the IP address of the user receiving this answer and (F_1, S_1) (F_2, S_2) ...(F_n, S_n) is a list of file identifiers matching the keywords, together with one provider for each file.
The second type of queries is of the following form:

$$T \;\; IP \;\; F_1 \;\; F_2 \;\; ... \;\; F_n,$$

where F_1 F_2 ... F_n is a list of file identifiers the user wants to download. The server's answers to these queries have the following form:

$$T \;\; IP \;\; (F_1, S_{11}...S_{1n_1}) \;\; (F_2, S_{21}...S_{2n_2}) \;\; ... \;\; (F_n, S_{n1}...S_{nn_n}),$$

where $S_{k1}...S_{kn_k}$ is a list of providers for file F_k.

The measurement lasted for 10 weeks, which represents 1 billion messages, with 89 million peers and 275 million files involved.

The second dataset consists in a capture of the logins and logouts of peers on an *eDonkey* server [8]. The login and logout information gives us the precise session length of users. A small number of session however present some problems:

- some sessions do not end in our dataset, most probably because the measurement stopped before the user disconnected;
- some sessions of a same user are nested within one another, for instance we observe two consecutive logins followed by two consecutive logouts. It is not possible in this case to know which logout corresponds to which login, and therefore we do not know the session length.

We discarded these two types of sessions in our analysis (they represent approximately 2% of all sessions). This dataset contains more than 200 millions of connections by more than 14 millions of peers, over a period of 27 days.

The two datasets are complementary: the first one does not give connection and disconnection times of users, and the second one does not contain information about queries. In the following, we will call the first dataset the *queries dataset* and the second one the *logins dataset*.

3. Users' session lengths – *queries dataset*

Here, we study the property S_k corresponding to the session length distribution, in the *queries dataset*. Since the session lengths are not directly available in this dataset (see Section 2.2), we have to infer them from the study of the queries made by a user. We detail this below, before turning to the actual study of the session length distribution.

3.1. Identification of users and sessions

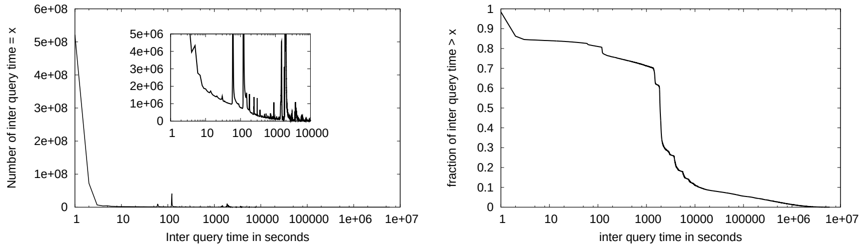
Identifying users in our data is a difficult question. We only have access to the IP addresses of the computers from which queries are entered. A computer is identified by an IP address at a given time, but this may change and we are unable in general to detect that a same computer has two different addresses (because of dynamic addresses for instance) and/or that two computers are using the same address (because they are behind a same NAT for instance). In addition, a same user may use several computers, and several users may use the same computer, making identification of users even more challenging. In the absence of a satisfying method for identifying users, there are two natural solutions: the first one consists in considering that a user corresponds to an IP address, and the other one consists in considering that a user corresponds to an IP address, together with the UDP port used.

We use here the first one, which allows to capture meaningful sessions (as explained below) and is therefore relevant. Moreover, we performed the same analysis by using the second definition, which also ensures the reliability of our results.

We infer sessions for a given user by studying the time elapsed between consecutive queries.

It is natural to consider that two consecutive queries made by a same user belong to the same session if the time elapsed between them is short, and belong to two different sessions if it is long. The question is then to find an appropriate threshold for distinguishing between these two cases. In order to give an answer to this question, we studied the inter-query time distribution, presented in Figure 1 (we display both the distribution (a) and the complementary cumulative distribution (b)).

We observe clear peaks at 60 seconds and at multiples of it (120, 240, 300, 900, . . .) in the distribution (they can be more clearly seen in the inset). These peaks indicate that, though users decide which queries to make and when they make them, there is a strong influence of the protocol on the observed data: most client applications automatically perform periodical



(a) Distribution.

(b) Complementary cumulative distribution.

Figure 1: Inter-query time distribution, for the *queries dataset*.

queries. Although these peaks become smaller after 1800 s, a zoom on the plot (not presented here) shows that they are clearly defined for values up till at least 20 000 seconds.

In order to smooth out the plot, we consider the complementary cumulative distribution (Figure 1 (b)). There is a high density of values between 1 000 and a value slightly smaller than 10 000 (the slope of the distribution is steep in this region). Such a high density indicates normal inter-query lengths within a session, and choosing a threshold in this region or before it would have little meaning. Therefore, we argue that the threshold must be at least as large as 10 000 seconds.

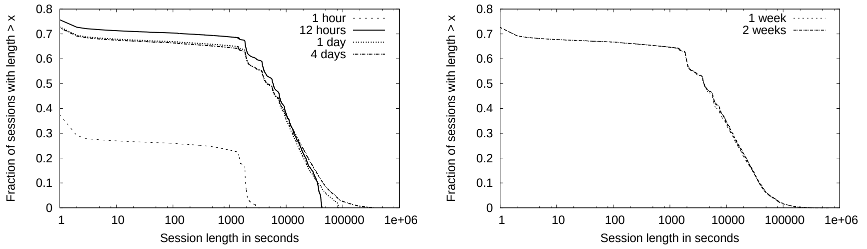
To study the importance of the peaks in the distribution, we computed, for a same measurement window, the session length distributions obtained with two different thresholds, the first chosen just before a peak and the second just after this same peak. We made a comparison between these two distributions and we observed no significant difference.

Finally, we have chosen to use a threshold of $t = 10\,800$ seconds, i.e. 3 hours. Therefore, in the following, if a same user sends consecutive queries separated by less than three hours, these queries belong to a same session, otherwise they belong to different sessions [2].

[2] A detailed study of session lengths would probably benefit from studying other values for this threshold. However our goal here is to illustrate our methodology and show that we can obtain interesting insights on the characteristics of session lengths. Other thresholds lead to similar results to this respect.

3.2. Characterization of session lengths

We now apply our methodology to the study of the session length distributions, by studying $S_k(W_{0,l})$ for different values of l.
We first observe that these distributions are highly irregular. They present clear peaks and valleys, which are linked to the peaks in the inter-session time distribution, see Figure 1 (a). Similar observations hold for different observation window lengths and positions. We will therefore consider complementary cumulative distributions, to smooth out the irregularities.



(a) l = 1h, 12h, 1 day and 4 days.

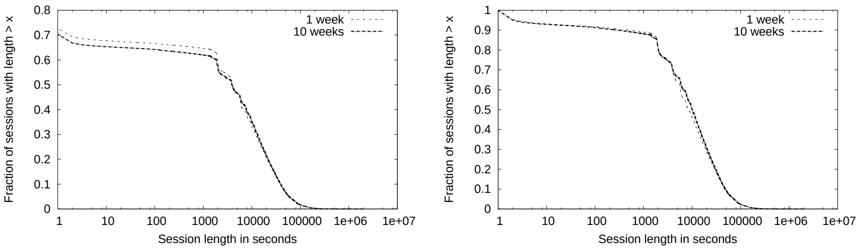
(b) l = 1 and 2 weeks (the two distributions overlap almost completely).

Figure 2: Complementary cumulative distributions of $S_k(W_{0,l})$ for different observation window lengths l, for the *queries dataset*.

Figure 2 presents the complementary cumulative distribution $S_k(W_{0,l})$ for different values of l, up to $l = 2$ weeks. The fractions of sessions with length 0 are not the same, which causes the normalized distributions to be vertically shifted[3]. The shapes of these distributions are however similar, with a small fraction of sessions with length smaller than 2 000 s, and an approximately linear shape between 2 000 s and 100 000 s. However, when $l \leq 1$ day, the distributions exhibit a clear cut-off. This is not the case anymore for $l \geq 4$ days: the tail of the distribution flattens after a bend occurring close to 100 000 s ($\sim$ 28 hours), and we observe a small fraction of *extreme* values after this bend. For observation windows larger than four days, the shape of the distribution does not seem to evolve anymore: Figure 2 (b) shows that the distributions for $l = 1$ week and $l = 2$ weeks are very similar to each other and to the one obtained for $l = 4$ days.

However, when l increases again, we observe a small difference between the corresponding distributions. Figure 3 (a) shows $S_k(W_{0,l})$ for $l = 1$ week

[3] Since the x-axis is in log-scale, the point $(0, 1)$ which belongs to all these distributions does not appear.



(a) Normalized distribution.

(b) Distribution normalized by number of values larger than 0.

Figure 3: Complementary cumulative distributions of $S_k(W_{0,l})$ for observation window lengths $l = 1$ week and $l = 10$ weeks, for the *queries dataset*.

and $l = 10$ weeks. We observe a small gap between them, caused by the fraction of sessions of length 0 (which again does not appear because of the log-scale on the x-axis): when the distribution is normalized by the number of sessions with length strictly larger than 0 (Figure 3 (b)), this gap disappears. This shows that, though the shape of the distribution does not vary anymore, the fraction of sessions with length 0 does.

When considering windows $W_{t_1,l}$ and $W_{t_2,l}$ of the same length but with different starting points, we observe that in general $S_0(W_{t_1,l}) \neq S_0(W_{t_2,l})$. As above, this difference is due to the fraction of sessions with length 0 which differs between these two distributions. This shows that the fraction $S_0(W_{t,l})$ of sessions with length 0 depends both on t and l, but that the general shape of the distribution, when this fraction is not taken into account, does not change.

We saw that the distributions seem *visually* not to change once the observation window length has reached four days. However, one must be careful when driving conclusions from a visual examination. Indeed, Figure 4 shows the distributions for $l = 1$ week and $l = 2 weeks$, but with a linear scale on the x-axis and a logarithmic scale on the y-axis. At first glance, the distributions seem strongly different from each other. However, a more careful examination shows that the distributions are similar for at least 99% of the values. They are different only for values larger than approximately 150 000 s, which

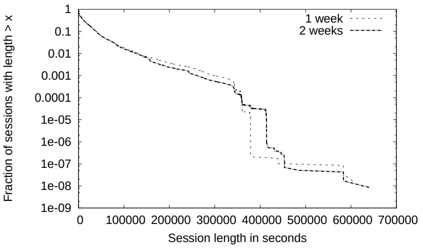
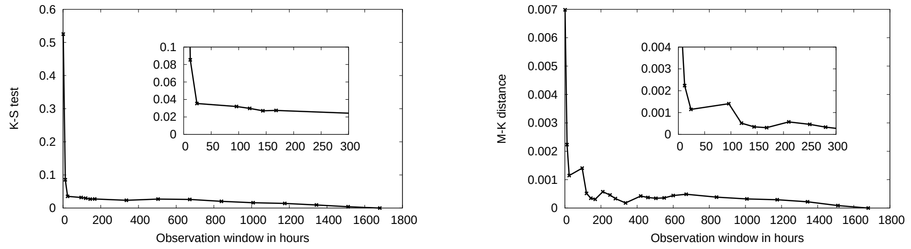


Figure 4: Complementary cumulative distributions of $S_k(W_{0,l})$ for observation window lengths $l = 1$ week and $l = 2$ weeks in lin-log scale, for the *queries dataset*.

are values seen after the bend of Figure 2 (b), and are significantly rarer than values below this bend. This leads us to consider them as *extreme* values. The fact that the extreme values change when l increases shows that they cannot be characterized with our methodology, and we leave their study for further work.



(a) $\mathrm{KS}(S_k(W_{0,l}),\ S_k(W_{0,l_{max}}))$ as a function of l.

(b) $\mathrm{MK}(S_k(W_{0,l}),\ S_k(W_{0,l_{max}}))$ as a function of l.

Figure 5: Study of the evolution of S_k with the K-S test and the M-K distance, for the *queries dataset*.

We now study the evolution of the distributions with the K-S test and M-K distance. Figure 5 (a) presents $\mathrm{KS}(S_k(W_{0,l}),\ S_k(W_{0,l_{max}}))$ as a function of l. The first values are high, and decrease quickly to approximately 4% for an observation window corresponding to $l = 24$ hours. After this, the decrease is linear. This clearly shows that observation windows of less than 24 hours are not representative. However, we do not know if the value 4% is small enough to consider that the distributions are similar or not. Moreover, the linear shape does not correspond to a value which fluctuates before becoming stable. This plot does therefore not allow us to decide when the observation window becomes long enough, or even to know if this happens during the measurement. Therefore, we cannot reach a conclusion with the K-S test.

We present the comparison with the M-K distance in Figure 5 (b): we compute $\mathrm{MK}(S(W_{0,l}),\ S(W_{0,l_{max}}))$ as a function of l. We observe a different behavior: the values observed tend to decrease (with fluctuations), until the observation window reaches approximately 150 hours (6 days and 6 hours). After this, the value of the M-K distance becomes very small: this shows that the corresponding distributions are very close to each other.

Finally, Figure 6 presents the standard deviation and the mean of $S_k(W_{0,l})$ as a function of l. We can see that the mean becomes stable once l reaches approximately 1 week, at the same time as the M-K distance. This confirms that an observation window of one week is long enough to accurately

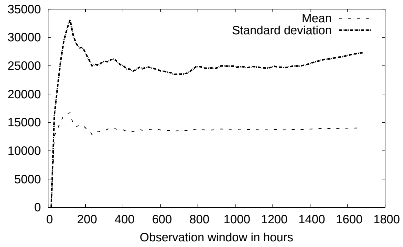


Figure 6: Mean and standard deviation of $S_k(W_{0,l})$, as a function of l, for the *queries dataset.*

estimate the distribution. The standard deviation, however, does not seem to stabilize as the observation window length increases[4], confirming that the distribution cannot be *fully* characterized. This is consistent with the distinction between the normal part of the distribution and extreme values. Indeed, the extreme values are very large and therefore have a strong impact on the standard deviation. The fact that they cannot be characterized causes the standard deviation to vary, whereas the fact that the normal part of the distribution is characterized causes the mean to become stable.

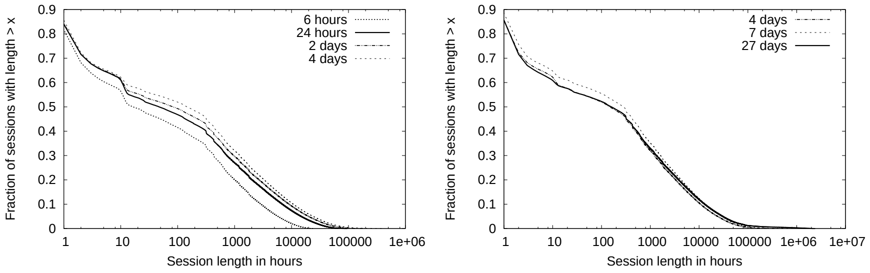
This confirms the intuition obtained by the visual study of the distributions: once the observation window length reaches one week, the normal part of the session length distribution stops evolving. This means two things. First, this distribution is stationary over time scales of the order of the whole measurement length, and it therefore makes sense to characterize it. Second, an observation window of one week is long enough to accurately estimate it. The extreme values of this distribution cannot however be characterized by our methodology.

4. Users' session lengths – *logins dataset*

We now study the session length distributions S_k, in the *logins dataset.*

Figure 7 shows the complementary cumulative distribution $S_k(W_{0,l})$ for different values of l, up to $l = 27$ days. We can see that the shape of these distributions are similar, and get closer to each other as l increases: in Figure 7 (a), we observe that the distribution corresponding to $l = 6$ hours

[4] Notice that, if we had stopped the measurement at 1200 hours, we would have had the impression that it stabilizes, hence the importance to have an observation window long enough.



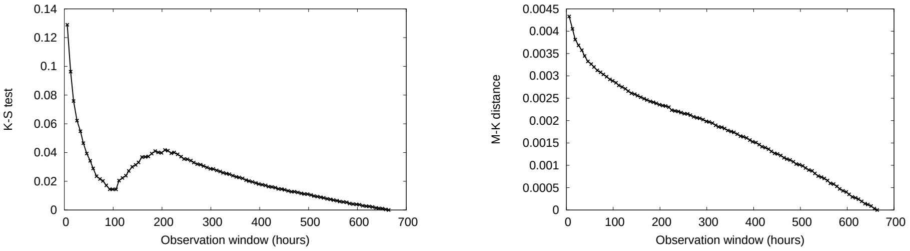
(a) $l = 6$ hours to 4 days

(b) $l = 4$ days to 27 days

Figure 7: Complementary cumulative distributions of $S_k(W_{0,l})$ for different observation window lengths l, for the *logins dataset*.

is a little different from the other distributions. For $l = 1$ day to $l = 4$ days, the distributions are closer to each other. When we increase l to 7 and 27 days (see Figure 7 (b)), the distributions remain close, but we observe also that the distribution corresponding to $l = 4$ days is closer to the distribution corresponding to $l = 27$ days than the one corresponding to $l = 7$ days.

In order to get a better intuition, we compare these distributions with the K-S test and M-K distance. Figure 8 (a) presents KS($S_k(W_{0,l})$, $S_k(W_{0,l_{max}})$) as a function of l. We can see that the values are high at the beginning, and decrease quickly to approximately 2% for an observation window corresponding to $l = 4$ days. After this, the values increase slightly until $l = 7$ days, which is consistent with our observations from Figure 7 (b). After $l = 200$ hours, the values tend to decrease almost linearly.



(a) KS($S_k(W_{0,l})$, $S_k(W_{0,l_{max}})$) as a function of l.

(b) MK($S_k(W_{0,l})$, $S_k(W_{0,l_{max}})$) as a function of l.

Figure 8: Study of the evolution of $S_k(W_{0,l})$ with the K-S test and the M-K distance, for the *logins dataset*.

When we compare the same distributions using the M-K distance (Figure 8 (b)), we do not observe the same phenomena. The values obtained tend to decrease linearly which means that the distributions change at a more or less constant rate. This shows that, though they are visually close and have a relatively small K-S test, the distributions corresponding to $l = 4$ days and $l = 27$ days are not this close to each other. Indeed, a more detailed examination of the distributions showed that the distance between them is not very large, but is present for a wide range for x values. The distance between the distributions corresponding to $l = 7$ days and $l = 27$ days is only large for small x values.

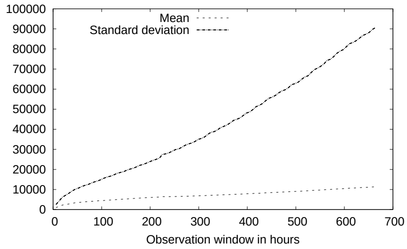


Figure 9: Mean and standard deviation of $S_k(W_{0,l})$, as a function of l, for the *logins dataset*.

Finally, we also compute the mean and the standard deviation of $S_k(W_{0,l})$ as a function of l, which we present in Figure 9. We observe that both of them increase linearly with the observation window length, which is consistent with the observations made with the M-K distance.

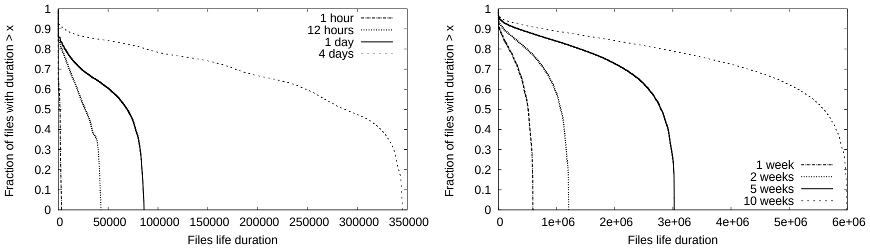
We have seen that the distributions $S_k(W_{0,l})$ are visually close to each other as soon as l is not too small. However, the numerical analysis shows that they evolve more or less linearly with l. Therefore, we cannot fully characterize this property because a longer or shorter measurement would exhibit a slightly different distribution. We however have confidence that the global *shape* of the distribution is the one we observed.

5. Files' life duration

We now study the files' life duration distribution, which we denote by F_k. Informations about files are only available in the *queries dataset*. In this section and the rest of this paper, we only consider the files for which there is at least one provider, because many files in the dataset are queried for but are never provided. These are files which don't exist in the system, at least during the measurement, and we therefore do not take them into account.

There are two possible ways to define a file's life duration. The first one is the same as for users' sessions lengths: considering that a file is not present in the system if there is no consecutive queries for this file distant from each other by less than a given threshold. In the second case, the life duration of a given file is defined by the time interval between the first and the last query for this file. Considering a threshold is not necessarily relevant here: we expect files to be more stable in the system than users, and the fact that a file is not queried for a (short) amount of time does not necessarily mean that it is not present in the system anymore.

We studied both definitions. In both cases this property does not stabilize. We present here the results obtained for the second definition, because they lead to interesting insight.



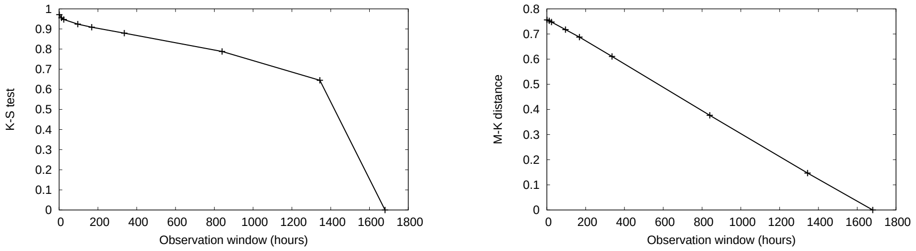
(a) $l = 1$ h, 12 h, 1 day and 4 days. (b) $l = 1$, 2, 5 and $l = 10$ weeks.

Figure 10: Complementary cumulative distributions of $F_k(W_{0,l})$ for different observation window lengths l.

Figure 10 (a) presents the complementary cumulative distribution $F_k(W_{0,l})$ for different values of l, from $l = 1$ hour to $l = 4$ days. We can see that the shape of the different distributions evolves strongly with l. This is also the case if we increase l and study the distributions corresponding to $l = 1$, 2, 5 and $l = 10$ weeks (Figure 10 (b)).

We observe that, the larger the observation window is, the larger the values of files' life durations tend to be: this can be explained by the fact that some files exist in the system for very long periods of time, so their observed life duration increases with the observation window length.

In order to confirm these observations more formally, we compare the distributions with the K-S test and M-K distance. Figure 11 (a) presents KS($F_k(W_{0,l})$, $F_k(W_{0,l_{max}})$) as a function of l. We can see that the values obtained are very high and vary much when l increases: for a measurement duration corresponding to $l = 1344$ hours (8 weeks), the K-S test value is still greater than 60%. We also compare the same distributions using the M-K distance and study MK($F_k(W_{0,l})$, $F_k(W_{0,l_{max}})$) as a function of l



(a) KS($F_k(W_{0,l}),\ F_k(W_{0,l_{max}})$) as a function of l.

(b) MK($F_k(W_{0,l}),\ F_k(W_{0,l_{max}})$) as a function of l.

Figure 11: Study of the evolution of $F_k(W_{0,l})$ with the K-S test and the M-K distance.

(Figure 11 (b)). It shows the same behavior as the K-S test: the values observed tend to decrease linearly and are very large.

We present in Figure 12 the mean and standard deviation of the distributions $F_k(W_{0,l})$, as a function of l. We can see that they both evolve continuously as the observation window length increases. These observations are consistent with Figure 10 and show that, the longer the observation window is, the larger the files' life durations are.

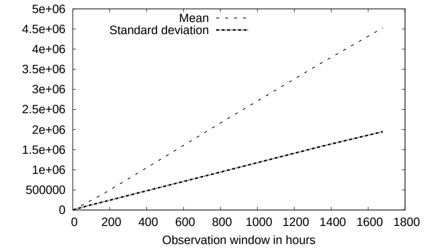


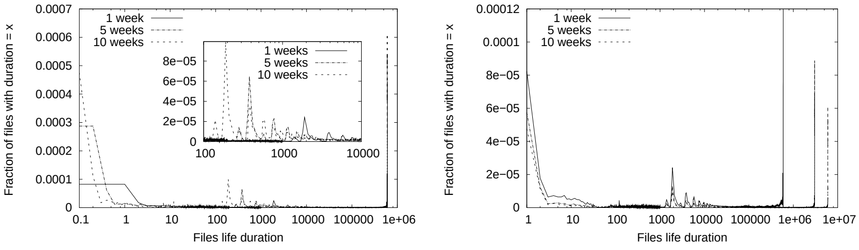
Figure 12: Mean and standard deviation of $F_k(W_{0,l})$, as a function of l.

We can investigate whether the distributions do evolve linearly with the observation window length l by normalizing them with respect to l. In order to do so, we divide the values of the x axis of the distribution $F_k(W_{0,l})$ by the observation window length l. To obtain normalized distributions, we also multiply the values of the y axis by l.

We present the corresponding normalized distributions in Figure 13 (a), for $l = 1, 5$ and 10 weeks. To better understand these plots, we also present the regular distributions (i.e., not normalized) in Figure 13 (b). We observe several things.

First, the normalized distributions all present peaks at the maximal possible values ($604,800s = 1$ week, which is the normalization unit for this plot). This correspond to the fact that a relatively large fraction of files have a life duration equal to the observation window length, as can be observed in Figure 13 (b).

Second, the normalized distributions present some intermediary peaks, which are not at the same x-values for the different distributions. This is caused by the fact that the regular distributions (Figure 13, b) present peaks which coincide. We observed a similar phenomena for users' session lengths in Section 3. This is caused by the fact that some clients send periodical queries, see Figure 1 (a). When the distributions are normalized, these peaks shift accordingly and the distributions do not coincide.



(a) Distributions normalized with respect to l.

(b) Regular Distributions.

Figure 13: Distributions of $F_k(W_{0,l})$ for observation window lengths $l = 1$, 5 and 10 weeks.

Since the K-S test and M-K distance can be computed only on cumulative distributions, it is not possible to compute them for the distributions shown in Figure 13 (there is no natural way to compute the cumulative of a distribution normalized in this way). We therefore just study the mean and standard deviation of the normalized distributions, which we present in Figure 14. We observe that, after some initial fluctuations, they both stabilize (note that the standard deviation stabilizes more quickly than the mean). It is interesting to note that the fact that the mean and standard deviation stabilize does not mean that the corresponding distributions also stabilize.

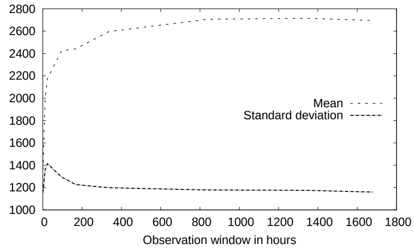
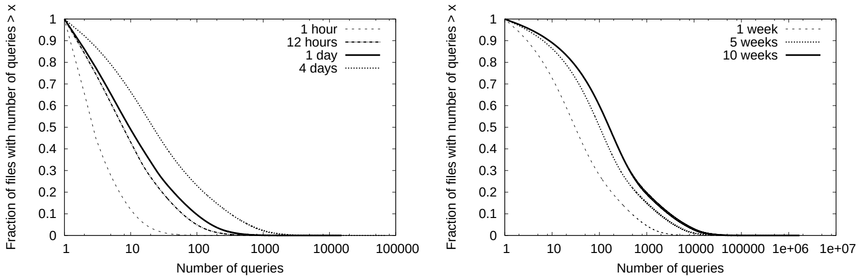


Figure 14: Mean and standard deviation (for distributions normalized with respect to l) of $F_k(W_{0,l})$, as a function of l.

Finally, this property cannot be characterized in our measurements. The (regular, un-normalized) distributions evolve continuously with the length of the observation window. Normalizing the distributions by the length of the observation window shows that this evolution is not regular, even though it is possible to characterize their mean and standard deviation. It remains an open question whether this property could be characterized if measurements longer than 10 weeks were performed, or whether it is intrinsically not stationary.

6. Number of queries per file

We now study the distribution of the number of queries per file Q_k, in the *queries dataset*.

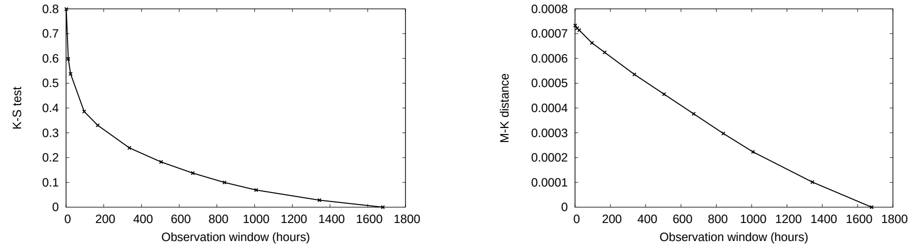


(a) $l = 1$ h, 12 h, 1 day and 4 days. (b) $l = 1$, 5 and $l = 10$ weeks.

Figure 15: Complementary cumulative distributions of $Q_k(W_{0,l})$ for different observation window lengths l.

Figure 15 presents the complementary cumulative distribution $Q_k(W_{0,l})$ for different values of l, from $l = 1$ hour to $l = 10$ weeks. We can see that the different distributions have some common properties: globally, we observe a linear shape at the beginning of each distribution which shows that there is a large fraction of files with a small number of queries (this fraction decreases as l increases). The tail of these distributions, however, tends to flatten which means that there is a small fraction of files with a very large number of queries. We observe also that the distributions evolve significantly with l: the number of queries per file increases with the observation window length.

We confirm that this property doesn't stabilize with the K-S test and M-K distance. Figure 16 (a) presents KS($Q_k(W_{0,l})$, $Q_k(W_{0,l_{max}})$) as a function of l. First, we can see that the values are very large: the values start almost at 80% for an observation window corresponding to $l = 12$ hours, to reach around 35% for $l = 1$ week. After this, the values tend to decrease linearly. The M-K distance follows almost the same behavior (Figure 16 (b)), except that the values tend to decrease more linearly. These observations are quite consistent with the ones obtained from Figure 15.



(a) KS($Q_k(W_{0,l})$, $Q_k(W_{0,l_{max}})$) as a function of l.

(b) MK($Q_k(W_{0,l})$, $Q_k(W_{0,l_{max}})$) as a function of l.

Figure 16: Study of the evolution of $Q_k(W_{0,l})$ with the K-S test and the M-K distance.

In Figure 17, we present the mean and the standard deviation of $Q_k(W_{0,l})$ as a function of l. We can see clearly that the values obtained for both, as expected, tend to increase linearly with the observation window length.

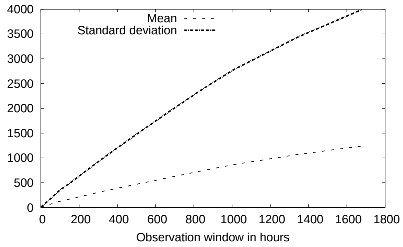
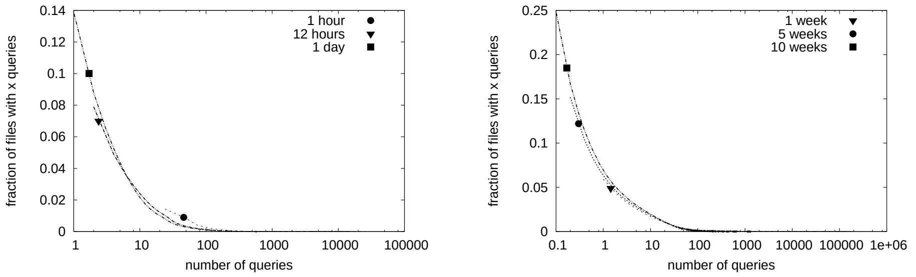


Figure 17: Mean and standard deviation of $Q_k(W_{0,l})$, as a function of l.

As seen before (for the file's life durations, see Section 5), the distributions seem to evolve linearly with the observation window length. In order to investigate this, we study the distributions normalized with respect to the observation window length. We perform this normalization in the same way as in the previous section, i.e. we divide the x values by l, and multiply the y values by l. The obtained distributions are shown in Figure 18. We can observe that the these distributions coincide, which means that they do evolve linearly with the observation window length.

This is confirmed by the mean and the standard deviation for the normalized distributions presented in Figure 19. We observe that the values obtained for the mean and the standard deviation follow the same behavior: at the beginning, they tend to decrease quickly, then stabilize once l reaches approximately 1 week. Note that the standard deviation decreases slightly with l, which seems to indicate that the proportion of very large values (af-



(a) $l = 1$ hour, 12 hours and 1 day.

(b) $l = 1$ week, 5 weeks and 10 weeks.

Figure 18: Distributions of $Q_k(W_{0,l})$ for different observation window lengths l, normalized with respect to the time duration.

ter the bend of Figure 15) tends to decrease. It remains an open question to see whether it would become completely stable with longer observation windows.

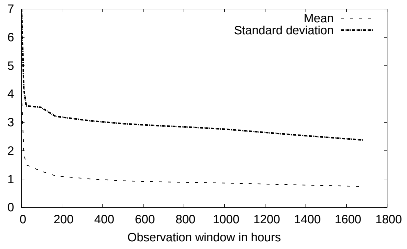


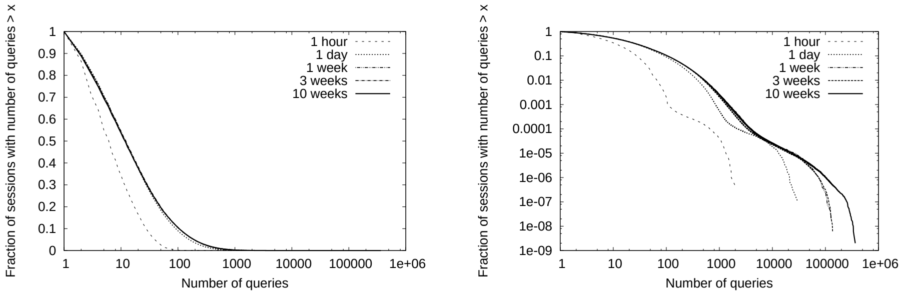
Figure 19: Mean and standard deviation (for distributions normalized with respect to the time duration) of $Q_k(W_{0,l})$, as a function of l.

Finally, we can conclude that the distributions of the number of queries per file evolve when the observation window length l increases. However, the study of these distributions normalized by l shows that this evolution is linear, which means that we are able to characterize this property.

7. Number of queries per session

We now study the distribution of the number of queries by session G_k, in the *queries dataset*. We consider the same definition of sessions as in Section 3.1, and study the number of queries the corresponding user performed within each session.

Figure 20 presents the complementary cumulative distribution $G_k(W_{0,l})$ for different values of l, from $l = 1$ hour to $l = 10$ weeks.



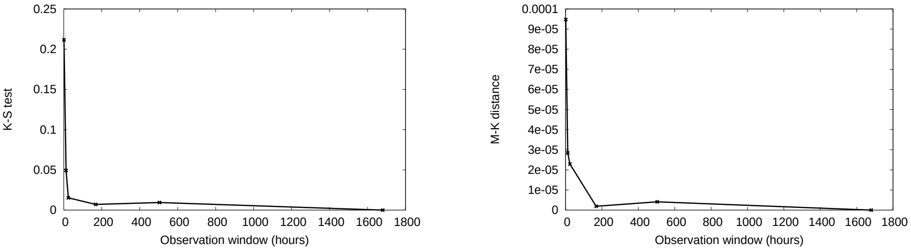
(a) log-lin scale.

(b) log-log scale.

Figure 20: Complementary cumulative distributions of $G_k(W_{0,l})$ for different observation windows lengths l.

In Figure 20 (a), we present these distributions in logarithmic scale on the x-axis and a linear scale on the y-axis. We can see that the shapes of the distributions are very similar, with a large fraction of sessions with a small number of queries and a small fraction of sessions with more than 1 000 queries. We observe that for an observation window larger than 1 day, the distributions overlap almost completely and do not seem to evolve when l increases.

When we compare the same distributions but with a logarithmic scale on both axis (Figure 20 (b)), we observe that they seem visually more different. However, we can observe that the distributions corresponding to $l = 1$, 3 and 10 weeks, are similar for more than 99% of the values. They are different only values larger than 1 000, which are after the bend of Figure 20 (a).



(a) KS($G_k(W_{0,l})$, $G_k(W_{0,l_{max}})$) as a function of l.

(b) MK($G_k(W_{0,l})$, $G_k(W_{0,l_{max}})$) as a function of l.

Figure 21: Study of the evolution of $G_k(W_{0,l})$ with the K-S test and the M-K distance.

Figure 21 presents KS($G_k(W_{0,l})$, $G_k(W_{0,l_{max}})$) and MK($G_k(W_{0,l})$, $G_k(W_{0,l_{max}})$) as a function of l. We can observe that they both follow the same behavior: the first values are high, and decrease quickly until $l = 24$ hours. After this, they decrease slightly and tend to stabilize after $l = 1$ week. This shows that the corresponding distributions are very close to each other which is quite consistent with our observations from Figure 20.

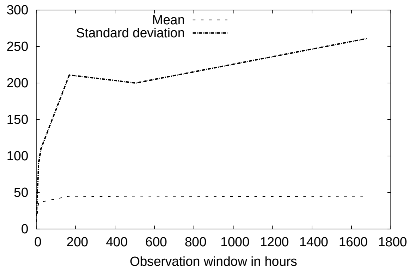


Figure 22: Mean and standard deviation of $G(W_{0,l})$, as a function of l.

In Figure 22, we present the mean and the standard deviation of $G_k(W_{0,l})$ as a function of l. We observe that the values of the mean decrease slightly at the beginning and become stable once l reaches 1 week, at the same time as the K-S test and the M-K distance. This shows that an observation window of one week is long enough to characterize the shape of this property. The standard deviation, however, does not seem to stabilize as the observation window length increases. This can be explained by the presence of very large values (larger than 1 000), which we have seen in Figure 20.

Finally, we can observe that this property has a very similar behavior with the first property we have studied (users' session length, Section 3). We distinguish two parts in the distribution: the first one corresponds to the large fraction of sessions with less than 1 000 queries, which we are able to characterize. The second one corresponds to the small fraction of extreme values which are not characterized by our methodology.

8. Related Work

The fact that the observation window length impacts the observed properties of a dynamic system has mainly been acknowledged for *churn*, i.e. the dynamicity of users, in P2P systems [2, 11, 12, 16, 13, 14].

Willinger *et al.* [17] addressed, in the context of IP flows, the question of whether the observation window is long enough to characterize some dynamic properties. They study the standard deviation of the flow size distribution as a function of the measurement length, and argue that the fact that it does not converge means that the samples may come from an underlying distribution with infinite variance. This in turn may make it difficult to fit the observed properties with a model.

The *create-based method* [11, 12] is based on the observation that being able to only capture accurately the length of sessions that begin and end within the measurement window creates a bias towards short sessions. To remove this bias, the measurement window of length T is divided into two halves, and only the sessions that begin during the first half and last less than $T/2$ are considered. This leads to an unbiased estimation of sessions with length less than $T/2$.

This methodology is complementary to the one we introduce here, which does not formally remove the bias, but allows to make observations for the shape of the distribution even for values larger than $T/2$. Moreover, our observations show that if the measurement window is too short, the create-based method will in some cases fail to provide an unbiased estimation. Finally, this method only applies to properties for which a notion of session can be defined, which is not always the case. For instance, it cannot be applied to the study of the number of queries for each file, which we performed in Section 6.

Finally, the bias caused by the finiteness of the observation window is not the only one occurring in our context. Stutzbach and Rejaie [13] studied different aspects of peer dynamics in three different classes of P2P systems (Gnutella, Kad and BitTorrent). They carefully analyzed the different kinds of bias that may influence such a study, and presented a list of those they identified, which includes problems linked to accurate peer identification.

Wang *et al.* [16] argue that the create-based method is biased when the data is obtained through periodic sampling, because short events may be missed or incorrectly observed. They propose a new sampling algorithm called RIDE (ResIDual-based Estimator) which measures session length distributions with high accuracy and requires a low sampling frequency.

Stutzbach *et al.* [14] investigate the issues arising when the whole system is not known, and informations about the nodes and links are obtained by a sampling procedure (in this case, random walk-based methods), in the case where the system evolves while the sampling process is under progress.

Friggeri *et al.* [4] studied contact networks captured with sensors able to detect when they are close to each other. They studied the bias on the observed contact duration caused by the fact that some sensors may fail to detect each other at some times.

9. Conclusion and Future Work

In this paper we introduced an empirical methodology for deciding when the bias induced by the finiteness of the observation window in dynamic systems becomes negligible. We illustrated its relevance by applying it to the study of several properties in a large P2P system.

This brought several key conclusions:

- if a system is observed for a period of time that is too short, it is not possible to obtain an accurate evaluation of its properties, which shows the relevance of our methodology;
- in a same system, it is possible to characterize some properties, but not others. This is the case for instance in the *queries dataset*, in which it is possible to characterize accurately the session length distribution, but not the file life duration distribution. This shows that there is no absolute relevant time scale to study a system, but that each property must be studied independently. This is confirmed by the fact that, for the properties that we were able to characterize, the minimum observation window length required is not exactly the same. Our methodology does not allow us to decide whether the properties that we were not able to characterize are not stationary, or if longer measurements would be required to characterize them;
- the degree to which we are able to characterize the system's properties varies: in some cases we are able to characterize the whole distribution, in others we can characterize the distribution except some extreme values, and in others we know the global shape of the distribution, but cannot trust its exact numerical properties. Knowing to which extent one can trust in a given property is a very valuable insight for the study of any system.

Finally, one key advantage of our methodology is that it can be applied to any property in any dynamic system, and allows to know which observed properties can be trusted and which cannot.

An interesting direction for extending this work would be to study models for the different properties we studied. This would allow us to gain a better intuition on the studied phenomena, and confirm formally our results. It may also provide *formal* bounds for the minimum observation window length needed to characterize a given property with a given accuracy.

Finally, we presented here a methodology for dealing with the bias introduced when measuring the dynamics of a system. In many systems, and in particular in the case of the internet, it is known that the measurement procedure may introduce a *structural* bias even if the system does not evolve with time. Some methods have been introduced to remedy this, see for instance [7, 6, 14]. We believe it is therefore crucial to combine methodologies such as the one we introduced here, which deal with the dynamic bias, to methodologies dealing with the structural bias, in order to capture the properties of systems such as the internet, as well as their dynamics.

Acknowledgments

We thank Pierre Borgnat, Patrice Abry, Matthieu Latapy, Marcelo Dias de Amorim and Nadjet Belblida for valuable ideas and comments about this manuscript. This work was supported in part by the French ANR MAPE project, and by a grant from the *Agence Nationale de la Recherche*, with reference ANR-10-JCJC-0202. This research work is conducted and funded by the European Commission through the EULER project (Grant No 258307) part of the Future Internet Research and Experimentation (FIRE) objective of the Seventh Framework Programme (FP7).

References

[1] Aidouni, F., Latapy, M., Magnien, C., 2009. Ten weeks in the life of an *eDonkey* server. In: Proceedings of HotP2P'09.

[2] Bhagwan, R., Savage, S., Voelker, G. M., 2003. Understanding availability. In: IPTPS. pp. 256–267.

[3] Chakravarti, I. M., Laha, R. G., Roy, J., 1967. Handbook of Methods of Applied Statistics. Vol. I. John Wiley and Sons, USE.

[4] Friggeri, A., Chelius, G., Fleury, E., Fraboulet, A., Mentré, F., Lucet, J.-C., 2011. Reconstructing social interactions using an unreliable wireless sensor network. Computer Communications 34 (5).

[5] Georgiou, T. T., Karlsson, J., Takyar, M. S., 2009. Metrics for power spectra: An axiomatic approach. IEEE Transactions on Signal Processing 57 (3), 859–867.

[6] Lakhina, A., Byers, J., Crovella, M., Xie, P., 2003. Sampling biases in IP topology measurements. In: IEEE INFOCOM.

[7] Latapy, M., Magnien, C., 2008. Complex network measurements: Estimating the relevance of observed properties. In: IEEE INFOCOM. IEEE, pp. 1660–1668.

[8] Le-Blond, S., Fessant, F. L., Merrer, E. L., 2009. Finding good partners in availability-aware P2P networks. In: SSS. pp. 472–484.

[9] Magnien, C., Ouédraogo, F., Valadon, G., Latapy, M., 2009. Fast dynamics in internet topology: Observations and first explanations. In: 2009 Fourth International Conference on Internet Monitoring and Protection. IEEE, pp. 137–142.

[10] Oliveira, R., Zhang, B., Zhang, L., 2007. Observing the evolution of internet AS topology. In: ACM SIGCOMM.

[11] Roselli, D., Lorch, J. R., Anderson, T. E., 2000. A comparison of file system workloads. In: Proc. of USENIX Annual Technical Conference.

[12] Saroiu, S., Gummadi, K. P., Gribble, S. D., 2003. Measuring and analyzing the characteristics of napster and gnutella hosts. Multimedia Systems 9, 170–184.

[13] Stutzbach, D., Rejaie, R., 2006. Understanding churn in peer-to-peer networks. In: Internet Measurement Conference. pp. 189–202.

[14] Stutzbach, D., Rejaie, R., Duffield, N., Sen, S., Willinger, W., 2009. On unbiased sampling for unstructured peer-to-peer networks. IEEE/ACM Transactions on Networking 17 (2).

[15] Torkjazi, M., Rejaie, R., Willinger, W., 2009. Hot today, gone tomorrow: On the migration of myspace users. In: Proceedings of the 2nd ACM SIGCOMM Workshop on Social Networks (WOSN'09).

[16] Wang, X., Yao, Z., Loguinov, D., 2007. Residual-based measurement of peer and link lifetimes in gnutella networks. In: INFOCOM. pp. 391–399.

[17] Willinger, W., Alderson, D., Li, L., 2004. A pragmatic approach to dealing with high-variability in network measurements. In: Internet Measurement Conference. pp. 88–100.