Revisiting pitfalls of DTN datasets statistical analysis

Gwilherm Baudic, Tanguy Pérennou, Emmanuel Lochin

Revisiting Pitfalls of DTN Datasets Statistical Analysis

Gwilherm Baudic
ISAE, Université de Toulouse
Toulouse, France
gwilherm.baudic@isae.fr

Tanguy Pérennou
ISAE, Université de Toulouse
Toulouse, France
tanguy.perennou@isae.fr

Emmanuel Lochin
ISAE, Université de Toulouse
Toulouse, France
emmanuel.lochin@isae.fr

ABSTRACT

Contact traces collected in real situations represent a popular material to assess the performance of a Delay Tolerant Network. These traces usually require some preprocessing to be fully usable. Especially, several assumptions can be made prior to performing the statistical analysis of contact and inter-contact times. We first classify these assumptions, and analyze their impact on the statistical characterization of three well-known datasets. We also identify some pitfalls in dataset analysis that might strongly influence the conclusion made by the experimenter. Based on our own experience, we subsequently propose a preliminary checklist to help researchers avoid undesired ambiguities or misunderstandings in further studies.

Categories and Subject Descriptors

C.2.1 [**Computer-Communication Networks**]: Network Architecture and Design—*Store and forward networks*

Keywords

DTN; Statistical Analysis; Real Traces

1. INTRODUCTION

With the growing use of mobile devices such as smartphones, Delay Tolerant Networks (DTNs) have gained a lot of attention. Assessing their performance largely relies on datasets of contact time (CT) and inter-contact time (ICT) between nodes, which are often generated from analytical models. To overcome their inherent lack of realism, datasets have been produced from field experiments [7, 3]. Because these are difficult to set up, only few research teams have managed to provide such datasets, leading to *de facto* standards. Subsequent investigations considerably benefited from the part of reality brought by these datasets. Statistical models have also been derived from them for CT and ICT distributions, either pairwise [2, 7] or aggregated [4].

Table 1: Main characteristics of the datasets

	Rollernet	MIT	Infocom '05
Duration (days)	0.125	284	3
Granularity (s)	15	300	120
Internal nodes	62	89	41
Internal contacts	60,146	114,046	22,459

To cope with the difficulty of data collection, the authors of EMO [6] propose a novel approach based on a simulation tool using node encounter events, which relies on CT and ICT distributions derived from datasets. Considering probability laws instead of raw data allows scaling of the studied networks both in time span and number of nodes.

However, the realism achieved with these datasets is questionable. A recent study shows that contact-based simulations ignore the limitations on node buffers and available transfer bandwidth [5], leading to biased estimations of delivery ratios and delays in DTNs.

In the present study, we will attempt to warn experimenters that the **statistical handling** of the datasets may alter them in a **somewhat hidden manner**. We focus on the statistical analysis of contact datasets and extract from previous literature on this topic a set of pre-analysis assumptions (Section 2). Then, we classify them according to their influence on the distribution fittings for three widely used datasets (Section 3). This allows us to propose a statistical analysis **checklist to avoid hidden pitfalls**, following our own experience (Section 4).

2. BACKGROUND

Contact traces have been widely used in the DTN literature as a basis for performance evaluation. Consequently, this section presents the traces exploited in this paper, along with the tools and assumptions used for statistical analyses.

2.1 Datasets used

We have selected three datasets for our study. They have already been widely used in the literature and are publicly available through the CRAWDAD archive website. An overview of their characteristics is presented in Table 1.

The first dataset is Rollernet [7], collected during a rollerskating tour in Paris in 2006. 62 Bluetooth contact loggers were distributed among nearly 2,500 participants. The second one comes from the Infocom 2005 experiment [3], which also used Bluetooth contact loggers, this time distributed among participants of the student workshop of the conference. Finally, the MIT Reality Mining dataset was collected

through an activity logging application embedded in mobile phones, lent to 100 students during 9 months of the 2004-2005 academic year. For this third dataset, we only considered the Bluetooth contact traces for the 89 devices which effectively recorded data. In all three experiments, devices performed periodic inquiry scans; this period is referred to as the *measurement granularity*. Since two nodes scanning simultaneously cannot see each other, a slight desynchronization of device clocks was voluntarily introduced.

Some limitations of the use of contact traces for simulations were already pointed out in the literature. The wireless technology used (e.g., Bluetooth) was previously shown to miss many contact opportunities, thus influencing subsequent performance results. In [5], the authors also emphasize the need to consider the real capacity of each contact opportunity, but this requires additional parameters not recorded in the datasets.

2.2 Fitting tools

Several approaches are used to fit the data to well-known statistical distributions. The simplest one is to plot the data with an adequate scale according to the model tested, and perform a graphical fitting. With this method, Hui et al. [3] fit the aggregated ICT in the Infocom 2005 dataset to a Pareto law. The authors of [4] later apply a similar approach to the same type of data (including the Infocom 2005 dataset), finding this time a Pareto law but with an exponential decay.

Other authors prefer to use statistical tools for their fitting studies: the work in [6] relies on the Kolmogorov-Smirnov (KS) test, when the authors of [2] and [7] choose the Cramer-von Mises test. While the latter can operate on discrete data, the former is restricted to continuous values and distributions. Considering that the recorded times in seconds are in fact discretized samples of a continuous variable, we choose the KS test for our studies.

2.3 Assumptions used in previous analyses

In this section, we list the choices and assumptions a practitioner has to make before a statistical analysis of contact traces. Based on previous articles, they are as follows:

Node choice: one can use only "internal" experiment nodes, or also include "external" nodes which were observed [3].

Symmetry of the pairs: a contact recorded between nodes i and j does not imply that j and i were also in contact, especially in Bluetooth traces. Thus, only a few papers [7] assume symmetry.

Minimum number of contacts: for pairwise metrics, a lower bound is usually chosen to guarantee enough samples for the statistical analysis: 9 contacts in [7], or 4 in [2].

0-second contacts: the traces used here all exhibit more than 40% of contacts lasting for 0 second. Yet, some papers appear to discard all contacts shorter than the measurement granularity [6], thus removing most of the data, while the authors of [7] include instead these 0-second contacts in their study by extending them to 1 second.

Time span: one might be tempted to use only a portion of a dataset. For the 284-day MIT dataset, 180 days[1] are considered in [6], while other authors use 246 days.

Inter-contact definition: although the common definition is the time interval when two nodes are not in contact, the difference between the *beginning* of two successive contacts is instead used in [6].

Power-law parameters: the Pareto distribution is almost always considered, as it has been shown in [3] to characterize the aggregated ICT and CT Complementary Cumulative Distribution Functions (CCDFs) of the Infocom 2005 dataset. In this case, the CCDF can be written:

$$F(x) = \left(\frac{x_{\min}}{x}\right)^{\alpha-1} \quad (1)$$

for $x \geq x_{\min} > 0$, $\alpha > 1$. The lower bound $x_{\min}$ is arbitrarily set as the measurement granularity in [6]. In [1], the authors propose a framework to replace graphical estimation techniques for *both* parameters (α and $x_{\min}$).

3. IMPACT OF INITIAL ASSUMPTIONS ON DATASET ANALYSES

We now present the consequences of the previous assumptions on the results of the statistical analysis. Unless otherwise stated, the baseline assumptions for this section consider asymmetrical pairs among internal nodes, with the inter-contact definition of Section 2.3. No pairs are discarded, and the measurement granularity is used as the Pareto lower bound $x_{\min}$. Finally, 0-second contacts are extended to 1 second, so that no inter-contacts are modified or removed. In the following subsections, we take each assumption separately while keeping the other parameters unchanged.

We chose to focus on the *aggregated* CTs, although our study also covered aggregated ICTs. We tried to fit them to three distributions: exponential, log-normal and power law (Pareto), which are the most represented in the literature.

All analyses were carried out using the R statistical software. We implemented maximum likelihood (ML) estimators to obtain the parameters of all the distributions considered, and used the companion code of [1] for the Pareto distribution functions and parameter estimation. Since traces record integer time values, ICTs and CTs typically exhibit ties (i.e., numerical values appearing more than once). Consequently, in order to use the correct empirical Cumulative Distribution Function (CDF) in the KS test, we reimplemented this test with the aid of the R `ecdf` function. Due to the presence of logarithms in the estimation formulas, 0-second contacts had to be extended to 1 second when taken into account, following [7].

We will present the effect of 0-second contacts, Pareto estimation methods, trace length and external nodes. The other hypotheses identified in Section 2.3 are not covered here, but are considered for future work.

3.1 0-second contacts

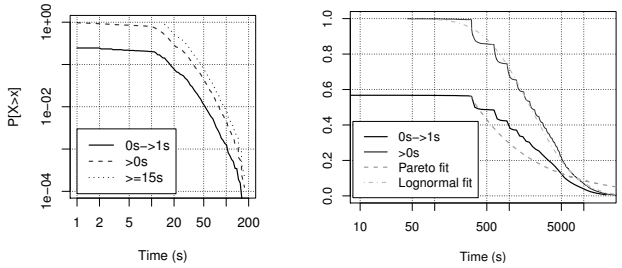
First, we study the influence of 0-second contacts, by investigating three possibilities: removal, and merging of the surrounding inter-contacts; extension to 1 second (baseline); removal of all contacts and inter-contacts shorter than the measurement granularity. We illustrate this with the 5000 first seconds of Rollernet, as in [7]. Note that this dataset assumes symmetrical pairs. The results are presented in Figure 1a for the aggregated CT distribution.

The three curves on Figure 1a exhibit large variations. The shortest contacts representing most of the trace (75% for 0-second contacts in this case), their removal also strongly

[1]Based on our findings, this corresponds to the 284 days with all weekends, MIT holidays and public holidays removed.

impacts the fitted parameters. For Rollernet, the best fits appear to be Pareto laws for all three hypotheses. Even with the same laws, parameters remain of utmost importance, as they can largely condition network behavior [4].

We now show the results for the 284-day version of the MIT trace in Figure 1b. For clarity, we did not represent the curve corresponding to the removal of all contacts shorter than the measurement granularity, as it would overlap the curve obtained with contacts strictly longer than 0 second.



(a) Rollernet (log-log scale) (b) MIT (log-lin scale)

Figure 1: Influence of the treatment of 0-second contacts on the aggregated CT distributions for two datasets.

As before, 0-second contacts create major changes on the curves. This time however, discarding these contacts also changes the best distribution from Pareto (with $\alpha = 1.534$ and $x_{\min} = 300$) to log-normal (with $\mu = 7.562$ and $\sigma = 1.130$).

When using traces to infer network capacity, 0-second contacts are paradoxical: 0 second means that no data can be transmitted, but the contacts do appear in the trace, implying indeed a data exchange. To solve this paradox, we believe that the extension to 1 second as in [7] is the best tradeoff between data transmission opportunity and short contact time. However, we just showed that this could greatly influence the statistical fittings. It is therefore crucial to check that 0-second contacts are treated consistently with the use cases of resulting models.

In this subsection, we applied only a lower bound to CTs, but upper bounds can also be set. The variability of the results exposed here, although restricted to a study on lower bounds, clearly calls for a better choice of filtering methods, as already mentioned in [2].

3.2 Pareto lower bound estimation

We now focus on the Pareto law. In Section 2.3, two methods for the lower bound estimation were presented: simply setting it to the measurement granularity, or using the algorithm introduced by [1] which selects the lower bound $x_{\min}$ and the associated exponent α as the ones providing the best fit, i.e., the smallest distance D in the KS test.

We compare these two methods for the aggregated CTs in the Infocom 2005 dataset. The results are shown in Figure 2. In this example, the distances D are 0.199 when the lower bound is set to the measurement granularity, i.e., $x_{\min} = 120$; and 0.027 with the estimated lower bound $x_{\min} = 1402$.

As can be seen from Figure 2, there are high discrepancies between the parameters provided by both approaches. We also found similar differences for the other datasets considered here, both for contacts and inter-contacts.

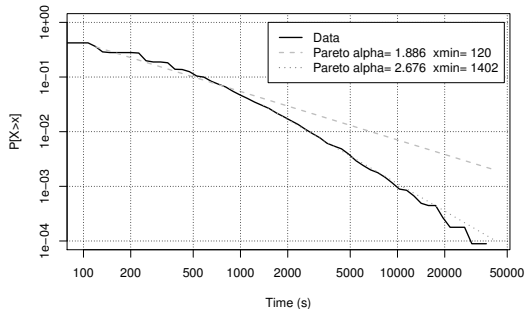


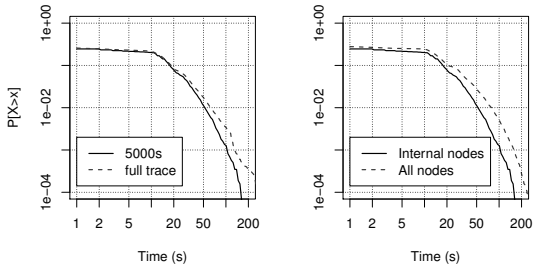
Figure 2: Pareto fittings of the aggregated CT distribution for the Infocom 2005 dataset (log-log scale).

In spite of being very contrasted, the results are not contradictory; they simply show that the lower bound giving the best fit (i.e., lowest KS test distance D) is not necessarily the measurement granularity, and is typically higher.

However, fitting the sole tail of the data is problematic, as it amounts to discarding most of the trace. In this case, the estimated lower bound $x_{\min}$ only covers 3% of the total data. With the measurement granularity, this percentage rises to almost 36%, still excluding a large part of the data. Consequently, some authors choose the measurement granularity as their lower bound and discard shorter contacts [6].

3.3 Trace length

Datasets are not always directly usable as a whole. For example, there may be periods when some of the nodes were not functioning properly, or the conditions changed (such as a break during a rollerskating tour [7], or holidays between school terms). Since these time periods may exhibit different properties, one might want to exclude them from the analysis. Hence, we analyze the influence of truncation using two datasets already studied in truncated versions: MIT and Rollernet. More precisely, we consider the MIT dataset in full (284 days) and without weekends and holidays (180 days) as in [6], while Rollernet will be used either in full, or only for the 5000 first seconds before the break.



(a) Trace length (b) External nodes

Figure 3: Influence of assumptions on the aggregated CT distribution for the Rollernet dataset (log-log scale).

The results are presented in Figure 3a for the aggregated CTs. For both examples, the change in duration only has a small impact on the curves. In the MIT case, omitted due to page constraints, the two curves even overlap each other. We also found similar results with the aggregated ICTs. The

case of the MIT dataset is particularly interesting: the 180 days correspond to only 63% of the total trace length, but account for 90% of the contacts recorded. This difference can be explained by the fact that the experiment was conducted among students, who are more likely to meet while on campus than during weekends or holidays. Due to the low variation in the total number of contacts between the two versions, similar results can be expected in both cases.

The two parts of the Rollernet dataset are much better balanced: first, the time percentage and the corresponding proportion of contacts are very close. Furthermore, these two parts were collected in similar conditions, except for the break when participants were likely to be less mobile. For these reasons, minor changes to the curves and the fitted parameters can also be expected.

3.4 External nodes

Traces are not restricted to contacts among the experimental devices; hence, one can choose whether to study or not these external contacts. However, as other devices may follow different mobility patterns, it may be interesting to assess their impact on the empirical distributions. We perform this evaluation for the 5000 first seconds of the Rollernet dataset. The results are shown in Figure 3b.

We find out that including the external nodes has a limited influence on the curves, for both CT and ICT empirical distributions. Our analysis of the Infocom dataset also leads to similar conclusions, as already found in [3]. Yet, the datasets capture rather different environments: a conference for Infocom and a rollerskating tour for Rollernet. While all attendants to a conference may have similar mobility patterns, passersby are typically slower and less mobile than roller-skaters. The number of external devices recorded is also very different between the two traces: 182 for Infocom and 1050 for Rollernet, much like the associated number of contacts: 5757 versus 43076. With such discrepancies on the percentage of external contacts, finding similar conclusions tends to indicate that external and internal nodes exhibit close behaviors, regardless of the experimental conditions.

4. CHECKLIST PROPOSAL

Sections 2 and 3 have described the initial assumptions usually made before statistical analysis of a CT/ICT dataset and their effect on the derived results. In this section, a checklist is proposed to keep the authors aware that apparently harmless assumptions might have a strong effect on subsequent results.

Did I use the whole dataset? Most authors filter the dataset, by discarding values (e.g., 0-second or very long contacts) or even periods (weekends, etc.) that do not fit the experiment they have in mind. This should be carefully described: what was filtered out? for what purpose?

Did I really *use the whole dataset?* The fitting method chosen might implicitly filter out a lot of data; for example, when setting the $x_{\min}$ threshold of the Pareto law. What happens to data samples lower than $x_{\min}$ should be stated: what data was left out? Should it be fitted in another way?

Did I change some values? Obviously, changing values will have an impact. However, some statistical fitting tools cannot be used with a dataset containing 0 values, such as those caused by the limitations of recording devices (e.g., Bluetooth scan time). With such values representing up to 70% of the contacts in Rollernet, the experimenter may have to explicitly choose between keeping the dataset, and not fitting Pareto or log-normal, or slightly changing it despite the above warning (e.g., extending 0-second contacts to 1 second).

5. CONCLUSION

In this paper, we studied the statistical analysis process of contact traces. First, we summarized the pre-analysis assumptions, based on previous works. Using three well-known datasets from the literature, we illustrated their influence on the parameters of the fitted probability distributions. We showed that 0-second contacts and the Pareto lower bound estimation have a strong impact, while trace length or external nodes play a smaller role.

Considering that accurate models need to be derived from real data, and that we previously showed preliminary assumptions can strongly affect these derivations, this reduces the field of use for such models: since a model would only capture one precise situation, it would be unsuitable for generalization. This represents another limit of contact datasets.

In fact, these conclusions highly depend on the nature of the datasets, and may not apply to other traces or some simulation/emulation setups. This motivates the checklist proposed here. This checklist is certainly not exhaustive, and should be expanded, at least with the other hypotheses mentioned in this paper. We would also like to extend our work to the pairwise metrics to compare with aggregated distributions.

6. ACKNOWLEDGMENTS

The authors would like to thank Aaron Clauset, Cosma Shalizi and Laurent Dubroca for the R code of [1], as well as Sebastien Ardon from NICTA for helpful discussions and the original EMO code [6].

7. REFERENCES

[1] A. Clauset, C. Shalizi, and M. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.

[2] V. Conan, J. Leguay, and T. Friedman. Characterizing pairwise inter-contact patterns in delay tolerant networks. In *Autonomics '07*, pages 1–9, 2007.

[3] P. Hui, A. Chaintreau, J. Scott, R. Gass, J. Crowcroft, and C. Diot. Pocket switched networks and human mobility in conference environments. In *Proc. SIGCOMM WDTN*, pages 244–251, 2005.

[4] T. Karagiannis, J.-Y. Le Boudec, and M. Vojnović. Power law and exponential decay of intercontact times between mobile devices. *IEEE Transactions on Mobile Computing*, 9(10):1377–1390, 2010.

[5] N. Ristanovic, G. Theodorakopoulos, and J.-Y. Le Boudec. Traps and pitfalls of using contact traces in performance studies of opportunistic networks. In *Proc. INFOCOM*, pages 1377–1385, 2012.

[6] F. Tan, Y. Borghol, and S. Ardon. EMO: A statistical encounter-based mobility model for simulating delay tolerant networks. In *Proc. WoWMoM*, pages 1–8, 2008.

[7] P. Tournoux, J. Leguay, F. Benbadis, V. Conan, M. Dias de Amorim, and J. Whitbeck. The accordion phenomenon: Analysis, characterization, and impact on DTN routing. In *Proc. INFOCOM*, pages 1116–1124, 2009.