# Towards a Scalable WiFi Monitoring System

Matteo Sammarco
UPMC Sorbonne Universités
matteo.sammarco@lip6.fr

Miguel Elias M. Campista
Universidade Federal do Rio de Janeiro
miguel@gta.ufrj.br

Marcelo Dias de Amorim
CNRS and UPMC Sorbonne Universités
marcelo.amorim@lip6.fr

## I. Context

Deploying a large-scale passive WLAN monitoring system raises an important issue: completeness. The trivial approach would increase the number of monitors until the merged trace becomes complete. The problem is the number of monitors required, which can drastically impact on the system scalability. The more traces to merge, the more CPU intensive becomes the whole procedure. *As a consequence, there is a clear tradeoff between completeness and scalability* [1].

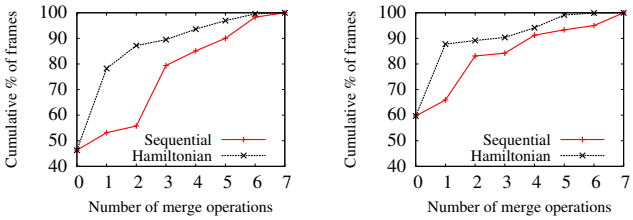
We propose a method called "Hamiltonian" to sort IEEE 802.11 traces before merging according to an ascending order of similarity. The first traces of the obtained sequence can contribute more to the merging procedure, whereas the last ones may be discarded to improve the system scalability.

## II. Merging Traces: Scalable Approach

We denote $\mathcal{T}$ the set of traces captured in the same time period, where $t_i$ is the trace collected by monitor $s_i$. We consider that each trace is composed of flows of frames, where $f_i^m$ the $m^{th}$ is the source-destination flow in trace $t_i$.

We weight the importance of the flow using the Term Frequency-Inverse Document Frequency ($\phi$) metric [2]. In our context, documents are traces and terms are flows. A vector of $F$ is assigned to each trace. The $n$ elements of the vector are the products of two factors: the flow frequency in that trace and the logarithm of the inverse trace (the one containing that flow) frequency over all the traces in the set. Hence, $\phi(t_i, n) = \frac{|f_n|}{|t_i|} \cdot \log \frac{|\mathcal{T}|}{|\{t_i \in \mathcal{T} | f_n \in t_i\}|}$. The similarity between traces $t_i$ and $t_j$ ($\sigma(t_i, t_j)$) is a value in the range $[0; 1]$ (from orthogonal to equal traces), given by the cosine of the angle between these vectors. This is equal to the dot product of the vectors, divided by the product of their magnitude: $\sigma(t_i, t_j) = \sum_n \phi(t_i, n) \cdot \phi(t_j, n) / [\sqrt{\sum_n \phi(t_i, n)^2} \cdot \sqrt{\sum_n \phi(t_j, n)^2}]$.

We consider the similarity values ($\sigma(t_i, t_j)$) between all the pairs of traces as the elements of the adjacency matrix of a fully connected graph $G(V, E)$. In this graph, each vertex $v_i$ corresponds to a captured trace $t_i$ and each edge $e_{ij}$ has a weight equals to the similarity value $\sigma(t_i, t_j)$. Our hypothesis is that touching all the nodes according to the minimum Hamiltonian path is a smarter way to iteratively select traces to merge because it ranks the traces according to their contribution to the final merge. This contribution can be estimated according to the weights of the edges connecting a trace to all the others. This rank is obtained from the path sequence, which is the solution of the Hamiltonian path problem. Merging a subset of top-ranked traces can improve the system scalability without losing representative information. We calculate the optimal Hamiltonian path with Concorde TSP Solver [3].




(a) IRCICA scenario (b) INRIA scenario.

Fig. 1. Comparison between Hamiltonian and sequential approaches.

## III. Evaluation

We have conducted experiments in two scenarios in Lille, France: IRCICA and INRIA. In both scenarios, we have deployed 8 monitors capturing Wi-Fi traffic during 100 minutes. Each monitor produces one trace at channel 1. We have adopted WiPal as a network sniffer and merging software [4].

We compare our proposal with a sequential strategy where traces are merged starting from monitor 1 then monitor 2 until monitor 8. For comparison purposes, we find the minimum Hamiltonian path starting from the trace captured by monitor 1 in both scenarios. In Figures 1(a) and 1(b), we observe how fast the Hamiltonian sequence converges to the total covering of unique captured frames. In two merging steps (or with the first three traces), traces chosen by the Hamiltonian sequence achieve more than 85% of the total unique captured frames in both cases. The same threshold is achieved by the sequential strategy only after a considerable higher number of traces. The Hamiltonian curve remains always above the sequential one, proving the importance of the right trace selection at the beginning of the whole merging process.

## Acknowledgment

The authors thank CNPq, FAPERJ, CAPES, FINEP, the SecFuNet project, and the ANR Rescue project.

## References

[1] K. Tan, C. McDonald, B. Vance, C. Arackaparambil, S. Bratus, and D. Kotz, "From MAP to DIST: the evolution of a large-scale WLAN monitoring system," *IEEE Transactions on Mobile Computing*, vol. PrePrints, no. 99, pp. 1–1, 2012.

[2] J. Ramos, "Using TF-IDF to determine word relevance in document queries," in *Instructional Conference on Machine Learning*, 2003.

[3] W. Cook, "Concorde TSP solver," http://www.tsp.gatech.edu/concorde.html, 2005.

[4] T. Claveirole and M. D. de Amorim, "Manipulating Wi-Fi packet traces with WiPal: design and experience," *Software Practice & Experience*, vol. 42, no. 5, pp. 585–599, May 2012.