Wireless Pers Commun (2016) 88:63–77
DOI 10.1007/s11277-016-3242-8

Data Driven Wireless Network Design: A Multi-level Modeling Approach

Carolina Fortuna[1] · Eli De Poorter[2] · Primož Škraba[1] · Ingrid Moerman[2]

Published online: 15 March 2016

Abstract Wireless network technology keeps improving by solving problems detected in current systems and anticipating requirements for future systems. One of the possible approaches to help advancing wireless technology is to develop methods that help researchers understand the less desired behaviors that may occur in a real-world system. One such method is data driven multi-level analysis that uses the monitoring data collected from real-world networks to provide detailed insight, at several levels and/or scales, into the system behavior. This paper discusses data driven multi-level analysis, provides a proof of concept on how it can be applied and identifies challenges. The contributions of this paper are (1) the use of data driven multi-level analysis for understanding the behaviour of wireless networks and (2) the identification of open challenges and directions for future research.

Keywords Wireless networks · Data driven research · Data science · Multi-level modeling

1 Introduction

Research areas as well as scientific approaches to investigate and understand wireless systems evolve over time. Initially, wireless research focused mainly on the analysis of single links or hops. With the introduction of ad-hoc networks, research focus shifted to include also multi-

✉ Carolina Fortuna
carolina.fortuna@ijs.si

Eli De Poorter
eli.depoorter@intec.ugent.be

Primož Škraba
primoz.skraba@ijs.si

Ingrid Moerman
ingrid.moerman@intec.ughent.be

[1] Jozef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia

[2] Ghent University - iMinds, Gaston Crommerlaan 8, 9000 Ghent, Belgium

hop networks, thereby increasing the complexity of the investigated systems. Although a limited number of military and academic experimental testbed facilities existed, most research was based on simulations or small manual ad-hoc deployments. As such, simulations were often used to analyse and model how such systems behave at scale.

As of around 2005, wireless testbed facilities have been developed and made accessible for research by the academic community. These testbed facilities enabled innovative large-scale research that included realistic behaviour of the devices and environmental conditions. These newly available testbeds spurred a growth of research focused on understanding and improving realistic system behaviour. As a result, the concept of "cognitive radios" and "cognitive networks" [24] became very popular around 2006. Cognitive solutions typically utilize a control loop to observe a system and its environment, in order to take appropriate actions to optimize the system, for example by opportunistically reusing available spectrum or by optimizing network parameters according to the dynamic wireless context. Existing cognitive solutions typically focus on solutions for self-optimizing and self-maintaining wireless systems, even in challenging but realistic conditions.

Rather than generate insight in the "how and why" a system exhibits a certain behaviour, cognitive solutions optimize a specific aspect of the network and as a result, multiple cognitive optimizations can even negatively influence each other [13]. To actually understand what is happening in a real and possibly large or dense wireless network, a number of recent scientific papers have urged to take a different "data driven" approach [4, 11]. They utilize large datasets containing experimental data to better understand the behaviour of wireless systems. These approaches (referred to as "data science" approaches) start from experimental data collected to provide insight into a specific behavior and utilize data mining techniques to better understand how the wireless system functions. Examples include the creation of system models [18] finding correlations and patterns in the considered parameter spaces, classifying and predicting outcomes or identifying trade-offs such as Pareto fronts. As such, research approaches have been shifting from optimizing realistic systems to the use of data science for understanding full system behavior in realistic conditions.

While data science typically focuses on explaining the most frequent behaviour of the system, it often overlooks seldom events such as a system crash. Multi-level (sometimes also referred to as multi-level/multi-scale) analysis uses models that are suitable for understanding instabilities (e.g. if there is an instability in a range, such as packet received ratio over distance, it will show up at multiple levels such as TCP window size, streamed video). Alternatively, multi-level/multi-scale allows us to understand trends in performance. As such, multi-level analysis provides the mathematical background to analyze such multidimensional inputs and to understand how these different scales (number of parameters, granularity, etc.) result in rare events (i.e. phase transitions). Percolation, a type of multi-level analysis has been recently investigated [7] in wireless networks, however only theoretical results using continuous variables are available while in this paper we use real data for the proof-of-concept multi-level analysis.

The contributions of this paper are (1) the use of data driven multi-level analysis for understanding the behaviour of wireless networks and (2) the identification of open challenges and directions for future research.

The paper is structured as follows. Section 2 explains what data driven research and multi-level analysis are. It also summarized the knowledge discovery methodology used for data driven research and discusses problems that could be solved using multi-level modeling. Section 3 presents a proof of concept of how to perform data driven multi-level analysis and identifies and future directions while Sect. 4 identifies existing challenges for such endeavors. Finally, Sect. 5 summarizes the paper.

2 Data Driven Research and Multi-level Analysis

2.1 Data Driven Research

Data driven research, nowadays often referred to as "data science", is research that puts a strong emphasis on starting from large data sets to solve a specific problem. This means that a relevant body of data has to be collected, processed, and understood. Traditionally, such tasks were performed by statisticians in collaboration with domain experts while now they are being performed by computer scientists working in the areas of data mining and machine learning. In order to perform data driven research in a specific field, also domain specific knowledge is needed besides computer science and statistical knowledge. Dhar [6] noticed that the "data" part of the science is important for "networks with complex relationships between their entities" such as the Internet, social networks, the Web and the emerging Internet of Things.

Wireless networks are an interesting application area for data science. They depend on electromagnetic propagation, which is a natural phenomenon, and on the network technology consisting of hardware and software elements that were build by humans. Modelling and understanding such systems at different scales requires us to distinguish human design imperfections and bugs from the underlying natural medium. However challenging, making this separation may enable the re-designing of systems based on the obtained knowledge.

2.1.1 Methodology

The ultimate goal of data science is to help discover new knowledge by looking into data. Data driven research has its roots in data mining and machine learning communities which have been traditionally concerned with handling large data, extracting knowledge and predicting or classifying new, previously unseen data (i.e. data which is not part of the training set). The de facto methodology for extracting knowledge from data relies on the *knowledge discovery process* which consists of 6 steps as depicted on the horizontal axis in Fig. 1.

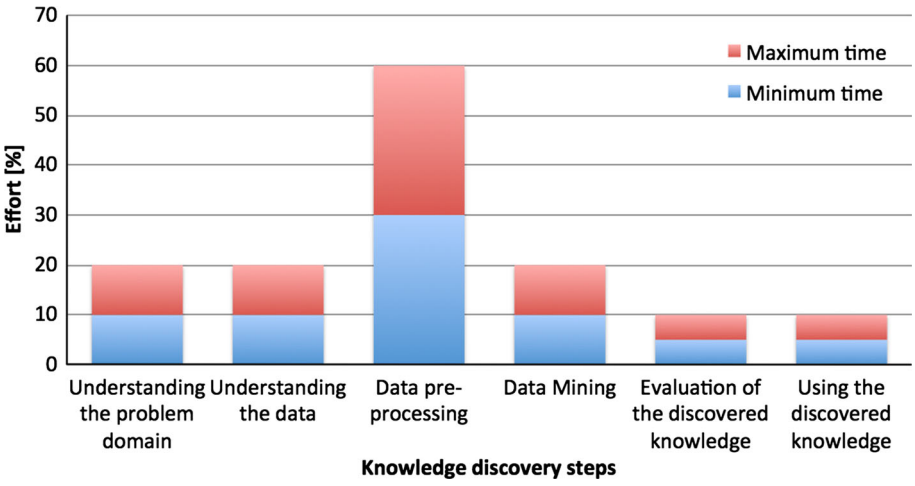


Fig. 1 The steps and min/max effort per step in the data driven knowledge discovery process (adapted from [14])

The first step of the knowledge discovery (KD) process is entitled *Understanding the problem domain* and is concerned with formulating the problem that can be solved using the data driven approach. The second step, entitled *Understanding the data* involves statistical analysis of the data to be used in solving the problem identified in the first step. The third step entitled *Data pre-processing* processes the data by cleaning it, fusing it, removing outliers and constructing features that will be then used by the machine learning or data mining algorithms. The fourth step entitled *Data mining* is concerned with training the data mining/machine learning algorithm with the data (feature vectors) obtained from the pre-processing step. The fifth step of the process evaluates the performance of the algorithms while the sixth and final step considers using the discovered knowledge in an actual implementation.

There are two main lessons to be learnt from the existing experience in knowledge discovery. First, *data* is at the centre of the process, if the data is not of good quality nor sufficient, knowledge cannot be extracted. Second, the data pre-processing step is always the most time consuming taking between 30 and 60 % of the total effort. This effort can be somewhat reduced if the degree of automation used in the research is increased as discussed later in Sect. 4.2.

2.2 Multilevel Modeling

One potential shortcoming of typical data driven approaches is the fact that they focus less on seldom events (e.g. sometimes referred to as anomalies). This practice is required to create approximations and models that explain most of the data (i.e. 90–95 %), but ignores the fact that less frequent behaviour (i.e. "unpredicted" behaviour) does occur and can have a significant impact on system behaviour [2, 20]—an intuitive example is the occurrence of a system crash. Multi-level models are able to recognize the existence of data hierarchies, which are characteristic for wireless networks. As such, they are appropriate to identify and explain the less frequently occurring phenomena [8, 20] (i.e. explain the heavy tails of the corresponding distributions—Fig. 2b).

A simple example of such infrequent phenomena is the occurrence of phase transitions (Fig. 2a). The functions on the figure can represent for example packet loss (y) in function of distance (X). While in the case of f1 there is a smooth transition between the parameters

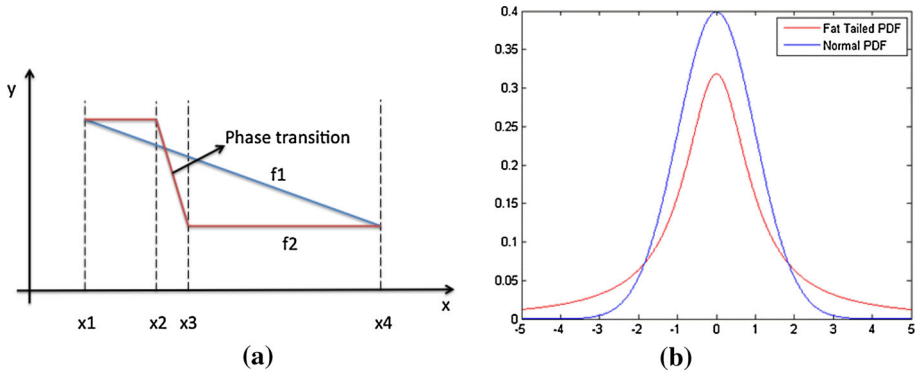


Fig. 2 Focus aspects for multilevel analysis. **a** Occurrence of phase transitions. **b** Heavy tail distribution due to phase transitions

between x1 and x4, in the case of f2 there is a phase shift between x2 and x3 (non-linearity or phase transition). Often such phase transitions are less frequent compared to the overall behaviour of the system and they manifest themselves through heavy-tail or power law distributions (as opposed to Gaussian distributions) where the side lobes converge to zero slowly (Fig. 2b). In many wireless behavioral patterns, stable operational regions are separated by phase transitions. These transition regions are often the causes of disputed or contradicting results as exemplified in:

- "The percentage of intermediate quality links was found significant in some empirical studies and insignificant in others" [2]
- "The claimed influence of ZigBee on WiFi ranges from insignificant to high impact" [2]

When considering complex systems, for example multi-dimensional wireless systems with parameters from multiple protocol layers, phase transitions can also occur in higher dimensional parameter spaces. These transitions are harder to visualize and typically require the use of advanced statistical theories such as multi-level modelling techniques to identify conditions (combination of parameters) that may have caused them. Multi-level modeling of complex dynamic systems is a recent research domain that can be used to enable the systematic study of the wireless networks and the complex interactions between the external and internal factors that govern their behavior. To this end, multi-level theory can take into account the following aspects:

- The scales that these studies consider vary and a behavior observed at one scale might not be noticed at a different scale [3, 8]. Examples of scales include: the granularity of the data, different network sizes, different network densities, etc.
- The analysis uses settings at a single layer—or from a limited set of network protocols—without fully understanding the interactions within the system, such as conflicting settings at different layers [13]. Whereas the examples above are mainly one-dimensional, phase shifts in complex systems are typically multi-dimensional. As such, for gaining insight in phase transitions in complex wireless systems, multi-dimensional inputs (with parameters from different protocols) should be considered.
- Studies that take into account the full complexity of wireless networks do not use experimental data but are mostly based on theoretical asymptotic analysis using simple models [8, 9].

3 Proof of Concept

In order to illustrate how the multi-level analysis enables powerful insights into wireless network behavior, we consider a ping response time dataset collected over three months from the LOG-a-TEC wireless testbed located outdoors [22]. The considered network has a simple star topology in which the coordinator periodically pings all the nodes. Each link is recorded at different time scales (i.e. averaged over): 300, 1800, 7200 and 86,400 s. Using the methodology specific to the knowledge discovery process summarized in Sect. 2.1.1, we present an example of performing multi-level analysis.

Understanding the problem domain: from links to networks Individual wireless links have been well studied even though some contradictions still exist as mentioned above and identified in [2]. Capturing network behavior, even for simple networks, is more difficult due to dependences between the different measurements (e.g. individual link qualities).

Furthermore, the temporal (i.e. time-varying) nature of the data further complicates interpretation and analysis, as does potential multi-scale behaviour. In this case multi-scale can refer to spatial effects such as an obstruction or source of interference which affects some part of the network (i.e. a tree) or temporal network-wide effects which are transient such as congestion.

One approach to characterizing wireless network behavior is to model a network as a space. This is generally not Euclidean space, but has some metric structure on it. This geometric approach has proved fruitful in many areas such as social networks, but also in sensor networks where each node is sample of a scalar function (e.g. temperature, light, etc).

There are several ways of modeling networks as spaces. Figure 3a shows a graph that models a wireless networks just by considering edges that correspond to the wireless nodes in the network while Fig. 3b models the network by considering directional wireless links. For the illustration of our approach, we model the star topology as directed links using the convention from Fig. 3b.

Understanding the data The available data can be explored and evaluated in many ways. We refer the interested reader to an early text on exploratory data analysis [25].

In our simple example with a star topology we measured the RSSI from the central server unicasting to the outer nodes in turns. This type of measurement is typical—we only have data during when transmissions take place. In Fig. 4, we see a binary map where the *x*-axis represents the different nodes, the *y*-axis represents time, the black color represents a sample while the white color represents a missing sample. The overall map shows highly irregular sampling. It can be seen that even though wireless network monitoring data should have values available each 300 s, there are many missing values (i.e. no ping response was received, therefore no RSSI value is available)—thus the result is irregular sampling from a data science perspective. In the subsequent analysis, we concentrate on the highlighted region where our data is most complete.

Within the well sampled region, we still have a notion of different time scales. RSSI measurements are essentially averages over a given time window. The choice of window size corresponds to a "time scale" of the measurement. If it is too short, noise dominates the measurements, while if it is too long, we average out the signal too much. The goal of our approach is to understand the structure of the data at different choices of this parameter. In principle, between the two extremes, we expect to see stable structures—where the measurements look qualitatively similar for a range of parameter choices.

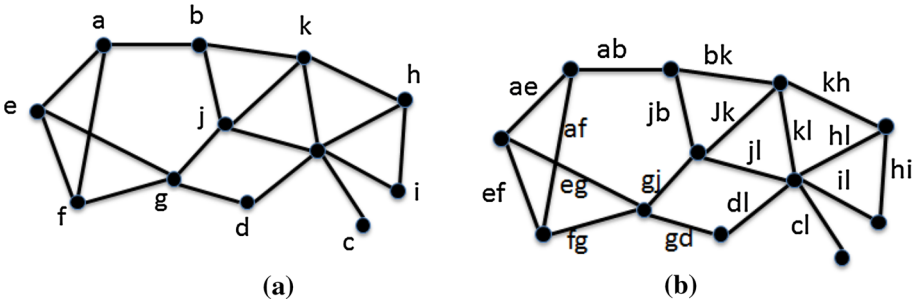


Fig. 3 Example network representation. **a** Example undirected graph. **b** The directed graph, the direction is given by the naming convention

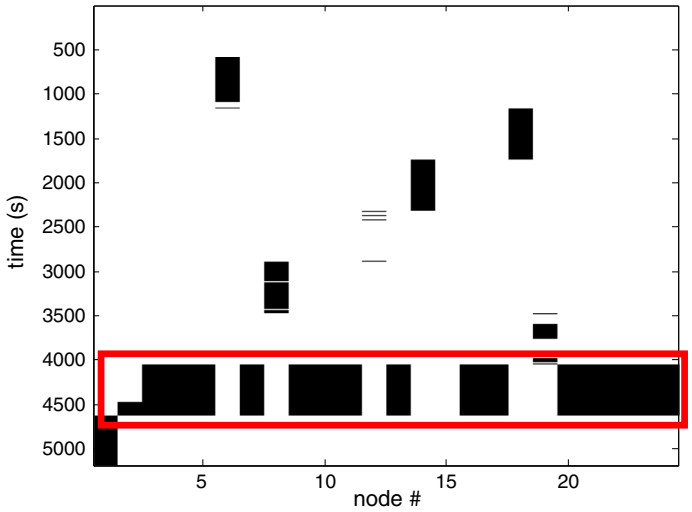


Fig. 4 Example wireless network monitoring data samples each 300 s. *Black* illustrates the existence of a sample, *white* the absence of samples. From a data science perspective, it presents highly irregular sampling as there are many missing values (i.e. no ping response was received, therefore no RSSI value is available)

Data pre-processing Our measurements are multi-variate since:

1. there are no obvious models beyond simple statistics (e.g. the probabilty of successfully sending a packet)
2. the random processes are highly non-Gaussian making closed-form solutions unlikely.

Non-parametric models are also not well suited to this problem since in high-dimensional settings such as this, they often require a large number of samples (roughly an exponential number with respect to the dimension).

A popular approach in machine learning and statistics to mitigating these problems is the use of *features* or *functionals*[1]. Functionals on spaces are used for generating features that are comparable and do not depend on the structure of the underlying graph (wireless network). Whereas a function takes in a point on a space and returns a scalar (Eq. 1), a functional takes in a *function* on a space and returns a scalar (Eq. 2). That is, a function is

$$f : X - > R \tag{1}$$

whereas a functional, takes a function and returns a scalar

$$F : f - > R \tag{2}$$

The prototypical example are kernels methods, bilinear functions which take pairs of points and return a scalar, etc. One common example is the radial basis function (RBF) kernel, defined as

$$K(x, x') = \exp\left(-\frac{d(x, x')^2}{2\sigma}\right)$$

Given two points, this function returns the negative exponential of the distance squared (normalized by a parameter σ). Due to its form, it is also known as the Gaussian kernel.

[1] The terminology depends on the field.

Using functionals we can generate feature vectors which allow for the use of machine learning, data mining, metric geometry algorithms to analyze the system in question. While we lose some information, the key obstacle in these techniques is generating relevant features (this is also referred to as feature engineering). This step is non-trivial, especially if we want to meaningfully compare the behavior of the underlying network at different scales or levels. For instance, how do we compare the behavior of a link with the behavior of the sub-graph containing all the 1-hop neighbors of the edges of the link? Or how do we compare the same network at different moments in time if one of the nodes has died and the entire topology changed as a result?

We illustrate that even very simple choices of functionals can shed light on the behavior of a system. In our simple example, we consider a feature vector which consists of the RSSIs sorted by strength. That is, we forget which link corresponds to which signal strength but rather treat all the signal strengths as a sorted list. This is illustrated in Fig. 5a. For the time interval in question, we visualize the function in Fig. 5b. This is not the only possibility nor do we claim any sort of optimality, but rather an illustration of the general approach.

Data mining The key advantage to this approach is that we can compute distances very efficiently between different points in time. Using the generated feature vectors, various data mining/machine learning algorithms such as clustering or multidimensional scaling can be used to study the network.

Treating the feature vectors at each time step as a point from Fig. 5b and applying multidimensional scaling (MDS) [21] to embed it into the plane we obtain the image shown in Fig. 6a. To measure the distance, we use Euclidean distance between the feature vectors (i.e. the sorted RSSI values). We cluster using single linkage clustering [19] to obtain the shown clusters (different clusters are shown in different colors). The corresponding dendrogram [12] visualizing the way the clusters of similar links are formed, can be seen in Fig. 6b. According to the results in Fig. 6a four clusters of RSSI values can be noticed, thus four types of links. Note that the axes in the figure are adimensional, only the

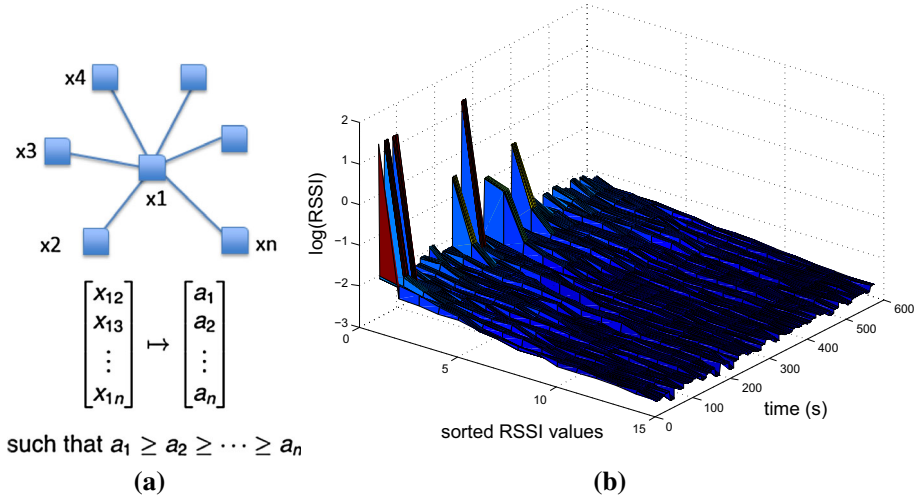


Fig. 5 Feature vectors. **a** Feature vector generation for a simple star topology and the RSSI levels. **b** Feature vector visualization

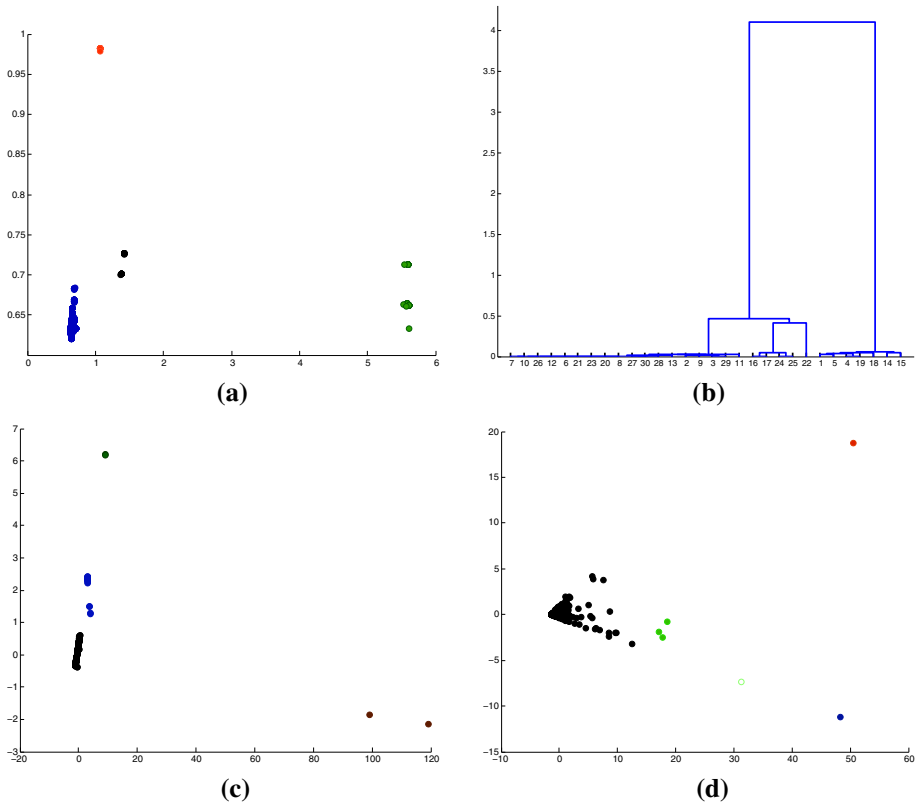


Fig. 6 MDS embeddings over different scales (window over which average is computed). The top right figure shows the dendrogram for single linkage clustering for the 300 averaging case. We can see that the clustering structure changes depending on the amount of averaging which is done to compute the RSSIs. **a** MDS embedding for averaging over 300 s (adimensional axes) **b** Dendrogram for averaging over 300 s (adimensional axes) **c** MDS embedding for averaging over 1800 s (adimensional axes) **d** MDS embedding for averaging over 7200 s (adimensional axes)

relative distance between the points is relevant, the closer the points, the more similar, also emphasized by different coloring.

The next step is to consider multiscale feature vectors. That is, different levels of averaging results in different functional values. As both averaging and embedding are continuous processes, we can compare different parameter choices. As can be seen in Fig. 6c, d, the structure of the functional changes depending on the scale parameter we choose. To interpret these results, we can look at the clusters in the time domain. We show the outlying clusters for each of the links for the first choice of scale in Fig. 7. From, here we see that the clusters are measuring a combination of very good links and bad links—with it being uninformative for some nodes. The four clusters are clearly distinguished by the distributions, however this illustrates the difficulty in interpreting the results directly (as well as finding these clusters by examining the time series directly).

The use of MDS and clustering is a basic illustration of the approach—indeed, one of the key advantages of using functionals, is that the result can be used in almost any machine learning algorithm, e.g. SVMs for classification, clustering based on similarities, anomaly detection, etc.

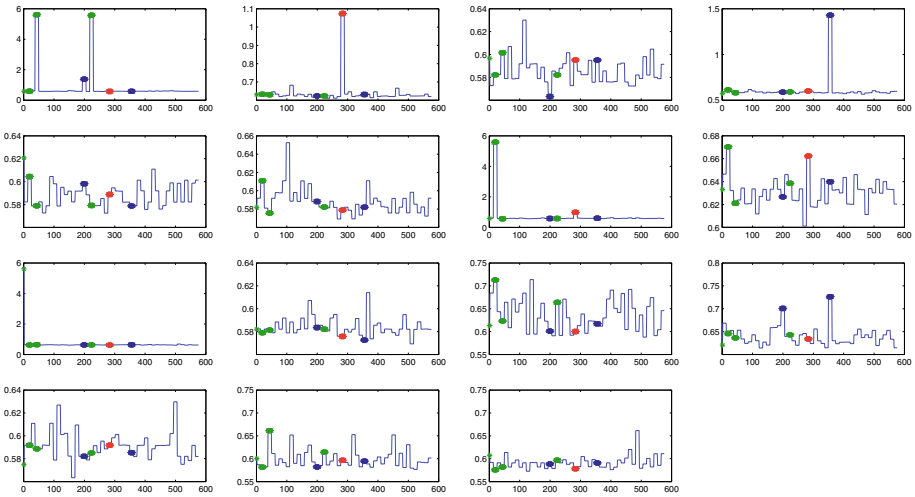


Fig. 7 The time steps for three of the clusters shown for the 15 nodes for which we have data over this time period

3.1 Future Directions

This is an illustrative example using a simple functional and basic data mining techniques. There are numerous directions from which to pursue this research. Immediately apparent is the use of more descriptive functionals—such as the full distributions (or higher moments) of the RSSIs values or functionals based on transforms such as the Fourier transform or wavelet-based transforms. Likewise, more advanced data mining techniques can perform tasks such as anomaly detection using wither annoyed examples or when more data is available (e.g. measurements over a longer period of time).

One of the most exciting possibilities coming from the multi-level modeling approach is the possibility of studying the dynamics of the collection of wireless channels. Takens' theorem [23], states that under mild conditions, given time series measurements of a system, it is possible to recover the dynamics by "raising the data" to a high enough dimension. The choice of the right functional should satisfy these requirements and therefore capture certain behaviors. It should be possible to recover periodic variations, such as daily variations as well characterize common disruptions, which may be recurrent rather than periodic. This may improve the prediction of wireless links by generating more accurate and robust models, with the benefit that since we are using functionals, we may use developed tools from machine learning and statistics (note that functionals often enjoy nice statistical properties such as Central Limit Theorems, exponential convergence, etc.).

This also provides new algorithmic questions, such as how to design network protocols which take advantage of this information as well as how to design online (and potentially distributed) techniques to compute and analyze functionals in-network rather than as an offline, global analysis.

4 Open Challenges

The main goal of data driven design is to speed up the improvements to wireless network design, preferably by providing a better understanding of heterogeneous and dense deployments at different scales. To this end, three main challenges need to be resolved by the community. The first one is related to realizing the supporting infrastructure needed to be able to experiment in heterogeneous environments of different scales. The second one is related to automating the collecting of the necessary raw data that can serve as an input to data driven models. The third challenges is related to extracting the meaningful knowledge for network designers from the raw data.

4.1 Infrastructure for Scaling Wireless Research

The first requirement for studying wireless networks at different scales is to have an infrastructure of such devices that would enable a systematic study. Existing infrastructures such as the ones available in ORBIT[2] and the Fed4FIRE federation[3], including the 10,000 node SmartSantander facility [10] are evolving towards supporting realistic, heterogeneous and dense set-ups with standardised APIs.

Even though most of these are predominantly homogeneous in terms of devices, capabilities and supported wireless technologies, many existing infrastructures are currently being extended to include heterogeneous devices and technologies. Similarly, although traditional facilities were often deployed in relatively artificial indoor environments, the number of outdoor deployments is also increasing. The methods to remotely access and control the testbeds are also improving by the switch towards tools that provide a uniform API rather than a testbed specific one. As such, it is now possible to automatically perform wireless experiments in multiple environments, using heterogeneous technologies and at different scales. The data about the wireless network has to be gathered from the devices themselves. This data includes the profile of the device in terms of capabilities, on-board events such as temperature increases and buffer overflows and data about links such as RSSI.

4.2 Increasing the Degree of Automation Used in Wireless Research

The degree of automation and "software-ization" in networking in general is increasing as can be seen by developments in Network Function Virtualization (NFV) and Software Defined Networking (SDN). Increasing the degree of automation with respect to (1) *controlling, configuring* and *upgrading* the elements of the network and (2) *data collection* from the networks are essential for maintaining and debugging the infrastructure. The more the testbeds are remotely accessible, controllable and configurable, the more they will be used for innovative research. Experimenters will run repeated experiments to solve the problem across different testbeds at various scales, times of the day and of the year, etc.

If the data collection is flexible and simple, it can be used to extract complex information and knowledge about the behaviour of the network at different scales. The data collection can be provided as an API where streams or measurements are delivered to the experimenter while the experiment is still running. An alternative is to store all the data and

[2] Open-Access Research Testbed for Next-Generation Wireless Networks (ORBIT), http://www.orbit-lab.org/.

[3] Federation for FIRE, http://www.fed4fire.eu/testbeds.html.

provide it to the experimenter once the experiment is finished. This data can also be stored and made publicly available so that other experimenters can use it at a later time for replicating results or for additional investigations. In both cases, it is important that the facility offers a wide flexibility in retrieving wireless network parameters. This translates into having an *easy to use, open* and *publicly accessible API* where the experimenter can easily extract all input and output parameters of both the devices and (testbed) environment.

It is important that the API is rich enough to provide enough *data* about the experiment, including the experiment descriptions and settings under which it has been performed—the so called *meta-data*. The API should also be extendable, meaning that additional observable parameters which are not necessarily directly connected to the experiment (i.e. MCU temperature, environmental parameters such as air temperature, etc.) could be added. To enable easy access to this data, the firmware /software should automatically be updated on the the devices from the wireless network.

Obvious evolutions with respect to increasing the degree of automation can already be observed. If we look at the work reported in [26], it can be seen that the experiments were collected on a local testbed to which the experimenters had physical access. In [16], the authors collect data from a local testbed and also from a remote testbed to which they have no physical access. Finally, in [17], the authors collect data from one local and two remote testbeds. This shows how, over time, the experimental infrastructures have evolved and it is now possible to do both data collection and experimentation remotely. The same conclusion probably will hold for the data processing tools that are used for the large traces that are collected to be analysed from the infrastructures.

4.3 Knowledge Extraction About Wireless Networks

Once a wide range of well-described experimental data is available through open and publicly accessible APIs, the next step is to extract knowledge from the data and use this knowledge for improving existing models or to design better systems. The data driven research methods suitable for extracting such knowledge can be adopted from more developed areas such as biological systems and the social web [11] where robust models explaining the observed patterns in complex dynamic systems already exist. For instance, the emergence of scales in very large data sets has been studied [3] with theoretical results also being available for wireless networks [1, 8]. For instance, in [8] scaling laws with respect to the throughput of wireless networks are derived, however they are yet to be empirically confirmed.

Knowledge extraction enabled by increasing the degree of automation in wireless research can have similar impact as the Web 2.0 has on social networks: large-scale social experiments can be preformed in days on several millions of users rather than requiring labor-intensive research taking months on a population orders of magnitude smaller. The most popular social networks include tens of millions of active users, all connected using web technologies and the information they generate is stored in *massive databases* and analysed using powerful data mining tools. The mining tools generate new knowledge about the network under investigation and, more recently, it is becoming easier and faster to *test* if the knowledge about a behaviour acquired in social science is correct. For instance, it is now feasible for a relatively small number of researchers in a relatively small time frame to investigate whether viral content can be predicted [5] and to actually engineer the desired viral content [15].

5 Summary

In this paper we proposed the use of data driven multi-level analysis for understanding the behaviour of wireless networks and identified open challenges and directions for future research. First, we explained what data driven research and multi-level analysis are. We summarized the knowledge discovery methodology used for data driven research as the same methodology is suitable for performing data driven multi-level analysis. We also discussed problems and contradictory results from recent literature that could be solved using multi-level modeling. In order to illustrate how data driven multi-level analysis can be performed, we provided a proof of concept using basic multi-level modeling methods on RSSI measurements from a simple wireless network. Future directions using more sophisticated multi-level methods on multidimensional data were identified. Existing challenges with respect to finding suitable infrastructures, collecting the data and performing the knowledge discovery using the proposed analysis have also been discussed.

Acknowledgments We would like to acknowledge Tomaz Šolc for sharing the data collected by the LOG-a-TEC wireless testbed monitoring system. This work was partly funded by the European Commission H2020 program under Grant Agreement Number 688116 (eWINE project), FP7 Grant Agreement Numbers 318493 (TOPOSYS project) and 612329 (Proasense project) and the IWT SBO SAMURAI project.

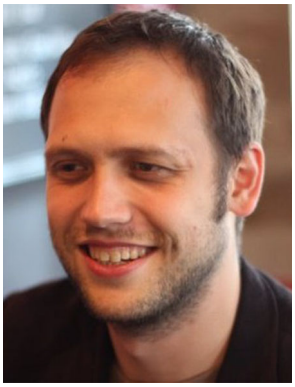
References

1. Ammari, H. M., & Das, S. K. (2008). Integrated coverage and connectivity in wireless sensor networks: A two-dimensional percolation problem. *IEEE Transactions on Computers*, *57*(10), 1423–1434.
2. Baccour, N., Koubaa, A., Mottola, L., Zuniga, M. A., Youssef, H., Boano, C. A., et al. (2012). Radio link quality estimation in wireless sensor networks: A survey. *ACM Transactions on Sensor Networks (TOSN)*, *8*(4), 34.
3. Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*(5439), 509–512.
4. Boano, C. A., Zúñiga, M., Brown, J., Roedig, U., Keppitiyagama, C., & Römer, K. (2014). Templab: A testbed infrastructure to study the impact of temperature on wireless sensor networks. In *Proceedings of the 13th international symposium on Information processing in sensor networks* (pp. 95–106). New York: IEEE Press.
5. Cheng, J., Adamic, L., Dow, P.A., Kleinberg, J.M., & Leskovec, J. (2014). Can cascades be predicted? In *Proceedings of the 23rd international conference on world wide web*, (pp. 925–936). International World Wide Web Conference Steering Committee.
6. Dhar, V. (2013). Data science and prediction. *Communications of the ACM*, *56*(12), 64–73.
7. Dousse, O., Franceschetti, M., Macris, N., Meester, R., & Thiran, P. (2006). Percolation in the signal to interference ratio graph. *Journal of Applied Probability, 43*(2), 552–562.
8. Dousse, O., Franceschetti, M., & Thiran, P. (2006). On the throughput scaling of wireless relay networks. *IEEE Transactions on Information Theory*, *52*(6), 2756–2761.
9. Franceschetti, M., & Meester, R. (2008). *Random networks for communication: From statistical physics to information systems* (Vol. 24). Cambridge: Cambridge University Press.
10. Gluhak, A., Krco, S., Nati, M., Pfisterer, D., Mitton, N., & Razafindralambo, T. (2011). A survey on facilities for experimental internet of things research. *IEEE on Communications Magazine*, *49*(11), 58–67.
11. González, M. C., & Barabási, A. L. (2007). Complex networks: From data to models. *Nature Physics*, *3*(4), 224–225.
12. Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, *32*(3), 241–254.
13. Kawadia, V., & Kumar, P. R. (2005). A cautionary perspective on cross-layer design. *IEEE on Wireless Communications*, *12*(1), 3–11.
14. Klösgen, W., & Zytkow, J.M. (2002). The knowledge discovery process. In W. Klösgen & J. Zytkow (Eds.), *Handbook of data mining and knowledge discovery*, (pp. 10–21). Oxford: Oxford University Press, Inc.

15. Lakkaraju, H., McAuley, J. J., & Leskovec, J. (2013). What's in a name? understanding the interplay between titles, content, and communities in social media. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*, Cambridge, Massachusetts, USA, July 8–11. Palo Alto, California: The AAAI Press.
16. Liu, T., & Cerpa, A. E. (2011) Foresee (4c): Wireless link prediction using link features. In *2011 10th International Conference on Information Processing in Sensor Networks (IPSN)*, (pp. 294–305). IEEE.
17. Liu, T., & Cerpa, A. E. (2014). Temporal adaptive link quality prediction with online learning. *ACM Transactions on Sensor Networks*, *10*(3), 1–41. doi:10.1145/2594766.
18. Mehari, M. T., De Poorter, E., Couckuyt, I., Deschrijver, D., Vanhie-Van Gerwen, J., Pareit, D., et al. (2015). Efficient global optimization of multi-parameter network problems on wireless testbeds. *Ad Hoc Networks*, *29*, 15–31.
19. Penrose, M. D. (1995). Single linkage clustering and continuum percolation. *Journal of Multivariate Analysis*, *53*(1), 94–109.
20. Ren, W., Zhao, Q., & Swami, A. (2014). Temporal traffic dynamics improve the connectivity of ad hoc cognitive radio networks. *IEEE/ACM Transactions on Networking (TON)*, *22*(1), 124–136.
21. Schiffman, S. S., Reynolds, M. L., Young, F. W., & Carroll, J. D. (1981). *Introduction to multidimensional scaling: Theory, methods, and applications*. New York: Academic Press.
22. Šolc, T., Fortuna, C., & Mohorčič, M. (2015) Low-cost testbed development and its applications in cognitive radio prototyping. In M.-G. Di Benedetto, A. Fabio Cattoni, J. Fiorina, F. Bader, & L. De Nardis (Eds.), *Cognitive radio and networking for heterogeneous wireless networks*, (pp. 361–405). Berlin: Springer.
23. Takens, F. (1981). *Detecting strange attractors in turbulence*. Berlin: Springer.
24. Thomas, R. W., Friend, D. H., DaSilva, L., Mackenzie, A. B., et al. (2006). Cognitive networks: Adaptation and learning to achieve end-to-end performance objectives. *IEEE on Communications Magazine*, *44*(12), 51–57.
25. Tukey, J. W. (1977). Exploratory data analysis. *Addison Wesley series in behavioral science: quantitative methods*. MA, United States: Addison Wesley.
26. Woo, A., Tong, T., & Culler, D. (2003) Taming the underlying challenges of reliable multihop routing in sensor networks. In *Proceedings of the 1st international conference on embedded networked sensor systems*, (pp. 14–27). New York: ACM.

Carolina Fortuna is currently research associate at the Department of Communication Systems at the Jozef Stefan Institute. She received her BSc in 2006, her Ph.D. in 2013 and was a postdoctoral research associate at IBCN, Ghent University 2014–2015. Her research is interdisciplinary focusing on semantic technologies with applications in modelling of communication and sensor systems and on combining semantic technologies, statistical learning and networks for analyzing large datasets. She has participated in H2020, FP7 and FP6 projects (H2020 eWINE, WiSHFUL, FP7 CITI-SENSE IP, FP7 CREW-IP, FP7 ACTIVE-IP, FP7-REGPOT-AgroSense, FP7 PlanetData-NoE, FP6-IST-CAPANINA); in CREW she was the technical leader of the JSI team. Carolina co-authored over 50 peer-reviewed publications and gained industry experience by interning with Bloomberg LP and Siemens PSE.

Eli De Poorter is a postdoctoral researcher at Ghent University. He received his master degree in Computer Science Engineering from Ghent University, Belgium, in 2006. He received his Ph.D. degree in 2011 at the Department of Information Technology at Ghent University through a Ph.D scholarship from the Institute for Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen). After obtaining his PhD, he received a FWO postdoctoral research grant and is now a post-doctoral fellow at the same research group, where he is currently involved in and/or research coordinator of several national and international projects. He is part of the program committee of several conferences and is the author or co-author of more than 70 papers published in international journals or in the proceedings of international conferences. His main research interests include wireless network protocols, network architectures, wireless sensor and ad hoc networks, future internet, self-learning networks and next-generation network architectures.



Primož Škraba received his Ph.D. in Electrical Engineering from Stanford University in 2009. He is currently at Researcher at the Jozef Stefan Institute in Slovenia and an assistant professor of computer science at the University of Primorska. His main research interests are applications of topology to computer science including data analysis, machine learning, sensor networks, and visualization. His other research interests include algorithms and optimization techniques.

Ingrid Moerman received her degree in Electrical Engineering (1987) and the Ph.D. degree (1992) from the Ghent University, where she became a part-time professor in 2000. She is a staff member of the research group on Internet-Based Communication Networks and Services, IBCN (www.ibcn.intec.ugent.be), where she is leading the research on mobile and wireless communication networks. Since 2006 she joined iMinds, where she is coordinating several interdisciplinary research projects. Her main research interests include: Sensor Networks, Cooperative and Cognitive Networks, Wireless Access, Self-Organizing Distributed Networks (Internet of Things) and Experimentally supported research. Ingrid Moerman has a longstanding experience in running national and EU research funded projects. At the European level, Ingrid Moerman is in particular very active in the FP7 FIRE (Future Internet Research and Experimentation) research area, where she is coordinating the CREW project and further participating in IP OpenLab, IP Fed4FIRE, STREP SPITFIRE, STREP EVARILOS, STREP FORGE and IP FLEX. In the FP7 research area on Future Networks, she is involved in IP LEXNET and STREP SEMAFOUR. Ingrid Moerman is author or co-author of more than 500 publications in international journals or conference proceedings. She is associate editor of the EURASIP Journal on Wireless Communications and Networking and vice-president of the expert panel on Informatics and Knowledge Technology of the Research Foundation Flanders (FWO).