# STRUCTURED REPRESENTATION OF NETWORKED DATA

*Santiago Segarra*

University of Pennsylvania, Dept. of Electrical and Systems Eng., Philadelphia, USA

As our lives become ever more integrated in the digital era, more of our actions and decisions are recorded, generating an unprecedented amount of data. This gathering of information leads to the emergence of massive datasets for which analytical tools are not yet well developed. Network data, i.e., data that encodes relationships between pairs of elements, belongs to this category. Networks, as ubiquitous structures for data representation, play a main role in our current scientific understanding of a wide range of disciplines ranging from biology to sociology. Although networks are not novel, an increasing interest in them was recently boosted partially due to the focus on distributed and decentralized algorithms.

Even though networks are widely used to represent data, there exist simple questions that we can ask when given a network for which only brute force answers are available. For example, if we are given a distance network and an element and are asked to search the network for the node that is closest to the given element, we have to compare the element against all the nodes in the network. Similarly, if we are given two networks and are asked how distant they are, we need to compare all possible permutations of the node sets. At a practical level, my thesis is concerned with finding answers to these questions that are smarter than brute force comparisons. At a foundational level, we believe that part of the difficulty in analyzing and efficiently managing large scale complex networks comes from the lack of structure that networks present. This loose nature contrasts with the rigidity of a closely related construction: the metric space. More specifically, if the network that we are given has a metric structure, we can find the closest node with a logarithmic number of pairwise comparisons and, while computing distances between metric spaces is still difficult, methods to compute approximated distances are available.

Given that understanding networks is challenging but understanding metric spaces is easier, a possible way of conducting network analysis is to project networks into metric spaces and then analyze the projected structures. The question that arises, then, is the design of methods and corresponding algorithms to implement these projections. Designing methods that enforce metric structure is not difficult. E.g., it suffices to replace the weight of each edge between two nodes by the minimum norm among all chains that link the given nodes. Using the 1-norm, this is equivalent to the shortest path distance between the adjacent nodes, but an infinite number of methods are possible since the choice of norm for the chain is arbitrary. It therefore seems that the important question is to decide which method is the most desirable out of the many ways to induce metric structure. Our approach to answering this question is axiomatic. More specifically, we deem as admissible a projection method that satisfies two axioms: a projection axiom – networks that are already metric remain unchanged after projection – and a dissimilarity reducing axiom – smaller networks have smaller metric projections. The apparent weakness of these axioms contrast with the stringent theoretical consequences. More formally, we showed that there is a unique admissible way of projecting weighted symmetric networks into metric spaces whereas for asymmetric networks there is an infinite but bounded family of admissible projection methods.

The key difference between metric spaces and weighted networks is that the former are governed by the triangle inequality. With this in mind, we extend the analysis of projection methods into metric spaces to projections into spaces governed by a generalized triangle inequality, which we call *dioid* metric spaces due to their close relation to the homonymous algebraic structure. Metric spaces are just one example of dioid metric spaces. Another remarkable example are ultrametric spaces, governed by the strong triangle inequality. Ultrametric spaces can be shown to be the output of hierarchical clustering methods. Thus, the same axioms conceived to develop admissible metric projections can be utilized to develop admissible hierarchical clustering methods. More general dioid metric spaces in which, e.g., distances need not be represented by scalars but can be elements of an arbitrary set, can be studied under this same framework. We can use this to represent a social network where every person has an opinion on a predefined set of issues and the distance between two people can be represented by a list of the issues on which they disagree. Although this network is very different from a metric space, both can be analyzed at the algebraic level of abstraction of dioids.

We exploit the fact that problems that are difficult to solve in networks with arbitrary structure become simpler once we project them into spaces with (dioid) metric structures to study two specific problems in network analysis: hierarchical clustering and search.

978-0-9928626-3-3/15/$31.00 ©2015 IEEE