Part I

Motivation and Gentle Introduction

Cambridge University Press
978-1-107-09611-0 — Wireless Physical Layer Network Coding
Jan Sykora , Alister Burr
Excerpt
More Information

1 Introduction

1.1 Introduction

Wireless networks are becoming more and more ubiquitous in the modern world, and more and more essential to today's society. In 30 years they have progressed from the province of a tiny minority of the world's population in only the most developed nations, to the point where there are very nearly as many wireless subscriptions as people in the world [24]. The services offered have extended from very limited speech services at the introduction of first-generation mobile systems in 1985, to broadband Internet access and full motion video today. Moreover, we are at the point where wireless networks will extend beyond connecting people (of whom there are a limited number), to connecting their devices – an effectively unlimited number. Some believe that there are already more devices than people connected to the Internet, and predictions that 50 billion or more devices will be connected by 2020 are circulating widely [60]. Of course, that is only the start.

All this implies that the *density* of wireless networks will inevitably increase. To provide telecommunication services to the human populations of our cities, at continually increasing data rates, will require increasing numbers of access points, for which backhaul will become an increasing problem, and require more widespread use of wireless backhaul. The devices will also form a network many times as dense as any current wireless networks, also likely to require connection to the core network. In both cases it is likely that the current point-to-multipoint architecture of wireless networks, exemplified by both cellular and WiFi systems, will be replaced by a multi-hop mesh network architecture.

The concept of the *mobile ad-hoc network* (MANET), one of the best-established concepts in wireless mesh networking, has been in existence for many years [9], yet has not really fulfilled its predicted potential. There are very few wireless networks in use today that implement a truly multi-hop networking approach. There seems to be a barrier to the practical implementation of multi-hop wireless networking that will surely have to be overcome in order to implement the ultra-dense wireless networks that are likely to be required in the near future.

Perhaps the most fundamental basis for such a barrier is that described by Gupta and Kumar in their well-known paper [20]. They show that for a conventional approach to wireless networking, in which transmissions from other nodes in the network are treated as interference, the total capacity of the network scales as the square root of the number of nodes – that is, the capacity per node decreases as the size of the network increases. Hence as networks become denser, and more hops are required, the capacity available to each terminal will decrease.

This interference problem has become widely recognized as the most significant problem limiting the performance of future wireless networks, including point-to-multipoint networks as well as multi-hop. Traditionally it has been mitigated by means of the cellular paradigm, which limits interference by ensuring that a certain *re-use distance* is respected. Increased density is accommodated by using smaller and smaller cells with greatly reduced transmit power, but this approach is now reaching its limit, both because of the large numbers of radio access points it requires and the resulting backhaul problem, and because cell sizes are becoming comparable in size with buildings and other city features.

All this suggests that it is time for a completely new paradigm in wireless networking, and a major objective of this book is to lay the foundations for such a paradigm, which we call the "Network-Aware Physical Layer."

1.2 The "Network-Aware Physical Layer"

Since the 1970s the design of communications networks has been based upon a *layered* paradigm, in which network functions are divided between protocol layers, each assumed to be transparent to the ones above it. The original layered model, dating from the late 1970s, was of course the OSI seven-layer model [2], but recently the layers implicitly defined in the TCP-IP protocol suite [1] have been more influential. In either case, the lower layers – the *network layer*, the *link layer*, and the *physical layer* – are of most interest to us here, since they provide the most basic functions of a communication network, namely routing, multiple access and error control, and modulation and coding, respectively.

Of these layers, the physical layer is the one that handles the signals which are actually transmitted over the communication medium: in our case these are related to the electromagnetic fields that form the radio waves. In the traditional layered paradigm the physical layer receives a signal from the medium and converts it to a bit stream, which is then passed to the link layer. However, this has the fundamental disadvantage that information is lost in the process that might improve the performance of functions which are located in higher layers. For example, it is well known that error correction is less efficient when operating on a bit stream (corresponding to *hard decision decoding*) than when it has access to a *soft decision metric*, which is usually obtained from the signal.

Moreover, it also means that signals from nodes other than the transmitter of interest must be treated as interference, which conveys no useful information but degrades the performance of the receiver in decoding the wanted signal. This arises because the traditional physical layer is assumed to operate on only one point-to-point link, which means signals on other links are interference (and vice versa). This is illustrated in Figure 1.1. The figure illustrates a multi-hop network in which data can travel from source to destination via two routes. We focus on the link of interest marked: in the traditional

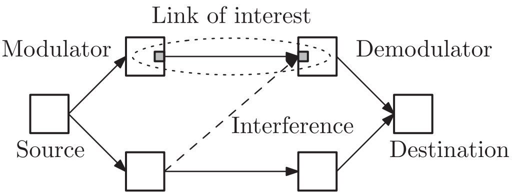

Figure 1.1 Traditional physical layer in a network.

paradigm the physical layer consists of the modulator at the transmitting node, the radio link between them, and the demodulator in the receiving node: that is, it relates to that link only, in isolation from the rest of the network. Thus a signal from another transmitter must be treated as interference (as shown), even though it carries information from the same original source, and could in principle be exploited to improve the reception of the data of interest.

Because interference is deleterious, it must usually be avoided wherever possible in traditional networks. This means that each node must transmit as far as possible on a channel orthogonal to the channel assigned to every other node – typically in a different time-slot or at a different frequency. This divides the resource available to the network and greatly reduces its efficiency. Again, information theory teaches us that greater capacity can often be achieved when multiple sources are allowed to transmit at the same time in non-orthogonal channels: for example, the capacity region of the multiple access channel (MAC) is achieved when the sources transmit simultaneously in the same channel, and is greater than the rate achieved by time-sharing of the channel.

The “network-aware” physical layer, on the other hand, does not need to nominate one node as transmitter of interest and hence treat all other signals but this one as interference. A network-aware receiver is aware – at the physical layer – of its location in the network, and what signals it may expect to receive in a given channel or time-slot. It is therefore able to determine what processing to apply to the composite signal formed by the mixture of all these signals. Similarly a network-aware transmitter is aware what effect its transmitted signals will have on other receivers, and can tailor the transmission in such a way that the received combination can also be processed as required.

Simply, if multiple interacting signals are unavoidable (e.g. due to the physical density of the network), it is better to make them useful one to each other as much as possible, instead of avoiding them. We do that directly on the signal level by properly constructing the transmitted coded signals and properly processing and decoding the received signals. This allows multiple nodes to transmit on the same channel, and avoids the division of resources. A receiver may even benefit from receiving combined signals rather than separate signals. It means that fewer signals have to be treated as deleterious interference, and any that do are typically weaker signals that have little effect.

This paradigm is not entirely novel: some functions which might be regarded as belonging to the link layer have already been implemented in the physical layer. One example is multiple access, which in computer networks is commonly implemented at the link layer by using protocols such as ALOHA or CSMA (Carrier Sense Multiple

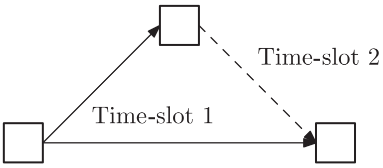

Figure 1.2 A simple cooperative communication system.

Access), or else is scheduled by using time-division or frequency-division multiple access (TDMA or FDMA). However code-division multiple access (CDMA), widely used in third-generation (3G) mobile systems, uses channels (corresponding to spreading codes) that are typically not fully orthogonal, and hence requires processing of the received mixed signal, which must be carried out at the physical layer, to separate the data. Similarly error control: while forward error correction (FEC) coding is conventionally regarded as part of the physical layer, retransmission protocols such as ARQ (Automatic Repeat reQuest) have traditionally been implemented at the link layer. However, recently hybrid FEC/ARQ schemes have somewhat blurred this distinction, since they require combining of signals transmitted in the course of multiple retransmissions.

Until recently, however, the functions of the network layer, specifically routing, have been excluded from the physical layer. This began to change about a decade ago with the introduction of *cooperative communications* [32]. Cooperative systems involve at least one relay node as well as the source and destination nodes (Figure 1.2), to assist the transmission of the source's data. Typically it receives the source signal in one time-slot, and retransmits it in some form in a subsequent slot. In most cases the processing within the relay is entirely at the physical layer, and frequently it is the original signal or some function of it that is retransmitted, without being converted to bits first. This is perhaps the simplest example of the physical layer being extended over a network involving multiple hops, beyond the simple link between one transmitter and one receiver.

This is, however, a very rudimentary version of routing. In this book we consider a much more general scenario involving multiple sources and multiple destinations, and multi-hop relaying between them. Thus routing is an essential element. The approach we will use, however, differs from routing in the conventional layered paradigm in two respects. The first is that it resembles cooperative communications in that processing within the relay takes place at the physical layer, involving signals directly. Unlike a bridge or a router in a conventional network, the relay does not decode the source data and transfer it to the link or network layer, but rather processes the received signals and forwards some function of them. The second is that what it forwards may not be a representation of data from a single source, but rather some function of data from several sources – a "mixture" of data from multiple sources to be separated at a later stage and delivered to the required destination. Thus it may no longer be possible to identify distinct routes for individual data streams, as is conventionally assumed.

This latter aspect can also be applied at the network layer of a multi-hop network, and corresponds to a technique introduced at the beginning of this century, known as *network coding*, which we will now discuss.

Cambridge University Press
978-1-107-09611-0 — Wireless Physical Layer Network Coding
Jan Sykora , Alister Burr
Excerpt
More Information

1.3 Network Coding at the Network Layer

Network layer *network coding* (NC) [5] addresses a network modeled as a directed graph connecting source nodes to destination nodes via a set of relaying nodes. In general there may be multiple sources and multiple destinations. The edges of the graph represent discrete links between pairs of nodes. This is clearly a good model of a data communications network with wired connections, such as the Internet, though we will see later that it does not represent a wireless network so well. For a unicast network, in which there is only one source and one destination, it can be proven that the maximum data flow rate is given by the *max-flow, min-cut theorem* [14]. However, Ahlswede *et al.* [5] showed that in the multicast case, where multiple destinations wish to receive the same data, the maximum flow rate cannot be achieved if relaying nodes operate simply as switches, connecting data flows on incoming links to outgoing links. Instead nodes should apply network coding, in which the symbols on an outgoing link are generated by some function of the symbols on two or more incoming links. This may be illustrated by the network shown in Figure 1.3, known as the *butterfly network*. The figure shows two versions of a network, in which two data sources each wish to send their data to both of two destinations, over a network in which all links have unit capacity. Figure 1.3a represents a conventional network in which the nodes can only switch a data stream from an incoming link onto an outgoing edge, or duplicate it and send on more than one outgoing edge. Thus the upper of the two relay nodes (which are marked as circles) can only select either stream A or stream B to send on its outgoing link (here it selects A). This is duplicated by the lower relay node, and hence the right-hand destination node can receive both streams, but the left-hand one receives only A. Figure 1.3b shows a network employing network coding. Here the upper relay node computes the exclusive OR (XOR) function (or modulo-2 sum) of the symbols in the data streams, and forwards the result. The lower relay node duplicates this to both destinations, and they can each recover both streams, because one is directly available, and the other can be reconstructed by reversing the network coding function applied at the relay node with the aid of the directly available stream. Thus the left-hand destination can now reconstruct stream B by applying $A \oplus (A \oplus B) = B$.

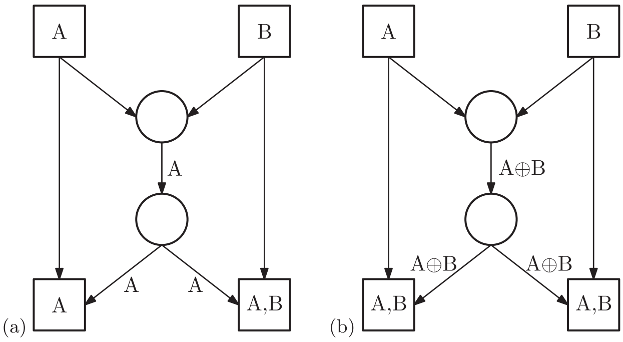

Figure 1.3 Butterfly network.

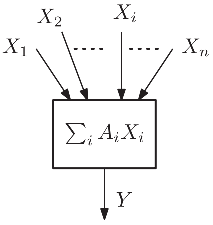

Figure 1.4 Linear network coding.

We will revisit this network topology later, in a slightly different context, but of course this principle also applies to much more complex networks, including networks containing cycles. Also in this case very simple coding is applied at the relay – simply the bit-by-bit XOR – but in general more complex encoding is required. There exists a wide variety of forms of coding, but [27] showed that linear coding over the finite field $\mathbb{F}_{2^m}$ is effective: in fact [34] had already shown that linear coding can achieve the maximum flow in a multicast network. Figure 1.4 illustrates this coding applied to a node: the output symbol Y is given by the formula in the diagram, in two different notations. In the first form $\otimes$ and $\oplus$ represent multiplication and addition within $\mathbb{F}_{2^m}$; in the second this is simply represented as a summation. The symbols on the incoming links are symbols in $\mathbb{F}_{2^m}$: they are drawn from an alphabet whose size is a power of 2, and can in fact be represented as length m binary strings. The coefficients A_i, $i = 1 \ldots n$ are also elements of $\mathbb{F}_{2^m}$, and again can be represented as length m binary strings. The addition operation is in fact simple bit-by-bit modulo-2 addition, but multiplication is more complicated: it is usually defined using primitive element operations on finite field (see Section A.2.1 or [8]).

It is clear that if all nodes apply a linear function of this sort, with symbols and coefficients from the same field, then the vector of output symbols across all relay nodes may be related to the vector of source symbols by a matrix. Equally clearly, for the destination nodes to reconstruct the source data this matrix must be full rank. We will revisit this model more rigorously later in the book.

1.4 Wireless Physical Layer Network Coding

The network model implicit in the conception of network coding, as illustrated in Figures 1.3 and 1.4, has one important deficiency as a representation of a wireless network. It assumes that the incoming links are *discrete*, and the symbols they carry are *separately* available to the network coding function in the node. This is a valid model of a wired network, but a wireless network does not have defined, discrete connections between nodes in the same way. Rather the electromagnetic fields due to signals transmitted simultaneously from two nodes will add together at the antenna of a receiving node, resulting in a superimposition of the two signals. Moreover they may be attenuated and/or phase shifted due to the wireless channel in largely unpredictable ways. In

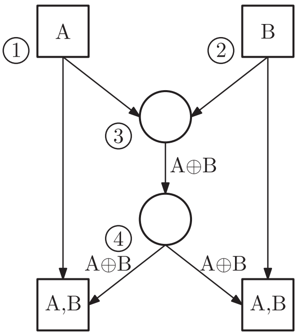

Figure 1.5 Network coded butterfly network with schedule.

the classical paradigm they are subject to fading and cause mutual interference to one another.

However, there are two approaches by which such discrete links can be emulated in a wireless network. The first is straightforward: separate orthogonal channels are provided for each link. In principle any means of orthogonalization could be used: different time-slots, different frequency channels, or different orthogonal bearer waveforms. For simplicity we will here assume that different time-slots are used: that the links are orthogonal in the time domain. Considering the network coded butterfly network in Figure 1.3b, this would require four time-slots per pair of source symbols to deliver the data to both destinations, as shown in Figure 1.5. This clearly reduces the efficiency of the network.

This also illustrates a general point about wireless networks that will be important in this book. Wireless devices are typically subject to the *half-duplex constraint*: that is, they cannot transmit and receive simultaneously on the same channel or in the same time-slot. There has been recent work on the implementation of full duplex wireless nodes, but that is beyond the scope of this book, in which for the most part we will assume the half-duplex constraint must be respected. This constraint immediately implies that a relay node can transmit in at most half of the time-slots.

As mentioned previously, information theory shows that transmission on orthogonal channels is not the optimum way of signaling from multiple source nodes to a single destination or relay node. In information theoretic terms this is known as the *multiple access channel* (MAC). The capacity of a MAC is defined by its *rate region*, as illustrated in Figure 1.6, for a two-user MAC. The left of the diagram illustrates the scenario: two sources, S_1 and S_2, transmit at rates R_1 and R_2 respectively to a common destination. The region within the solid line in the graph on the right denotes the rate region: the set of rate pairs that can be achieved with low error rate. Note that it implies that three limits operate: a limit on the rates R_1 and R_2 that each source can transmit independently plus a limit on the *sum rate* $R_1 + R_2$.

Note, however, that a conventional system using TDMA (i.e. using orthogonal time-slots) would be restricted to the triangular region shown by the dashed line — since any increase in the rate from one source would always have to be exactly balanced by

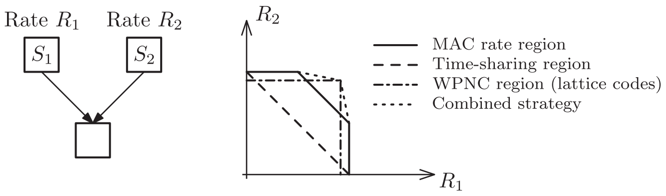

Figure 1.6 Rate region for two-user MAC.

a reduction in the rate from the other. The system can do better than time-sharing by allowing both sources to transmit simultaneously, and at the receiver to first decode one, then cancel the interference it causes and decode the other. This allows an increase in the sum rate significantly above the time-sharing rate. Thus in the network coded butterfly network we could allow sources A and B to transmit simultaneously, merging time slots 1 and 2 in the schedule shown in Figure 1.5, and increasing the network throughput.

However, this still constitutes a bottleneck in the network, because it requires symbols from both sources to be decoded even though what is required is only the one symbol formed by combining them with the network code function. Taking this into account, it is possible (as we will see later) to establish what we will call the *WPNC region*, which is the set of source rates which allows this symbol to be decoded. This is shown by the dash-dotted lines in Figure 1.6, and allows rates outside the conventional two-user MAC region. It is achievable e.g. by the use of nested lattice codes, as will be discussed in Chapter 5.

To achieve a rate outside the MAC region requires that rather than being obtained by decoding the two sources separately, and then applying network coding at the network layer (a strategy we will call *joint decoding*), the network coded symbol must be decoded directly from the received signal at the physical layer – in other words by *physical layer network coding* (PLNC). In this book we refer to the technique as *wireless physical layer network coding* (WPNC), and it is the main topic of the book. The term “wireless” is used here because the inherent superposition of wireless signals mentioned above means that this form of network coding is essential in wireless systems to obtain all the information available. There will of course be much more detail to come, and in particular there will be a “gentle” introduction to the main principles in the next chapter, so here we will restrict ourselves to a very simple example of how this might work and how it can enhance capacity.

Figure 1.7 shows the scenario. Two terminals transmit uncoded BPSK, taking signal values ± 1 over channels with the same attenuation and phase shift to a relay. We assume that the relay applies network coding using the XOR function. At the relay the signals add, resulting in the values ± 2 and 0. A joint detection strategy would need to decode the two sources separately, and this is clearly not possible if the value 0 is received, since it might represent the data 01 or 10. WPNC, on the other hand, has only to detect which network coded symbol the received signal corresponds to. This avoids the problem, since 01 and 10 both correspond to the network coded symbol 1. Thus the received signal can be interpreted as a constellation in which both the signals marked with white circles

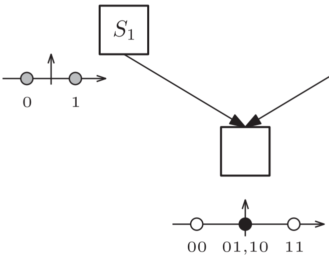

Figure 1.7 Illustration of PNC operation.

correspond to (network coded) 0, while the black circle corresponds to 1. This clearly increases capacity compared to both the joint decoding approach and the network coding approach.

1.5 Historical Perspective

At this point we will take a break from the technical details of WPNC to discuss how we reached this point, and the initial development of WPNC up to the present. We have already discussed some of the information theoretic background, and have mentioned the development of network coding. It is worth noting, however, that many of the theoretical foundations of multi-user information theory were laid in the 1970s – including analysis of the multiple access channel [4], [35], of the broadcast channel [11], and of the relay channel [12]. However, there has been little practical implementation of these concepts even up to today, although that is now changing, notably because of the pressures on wireless networks noted above, and also because multiple antenna systems have important synergies with the MAC and broadcast channels, which have led to the introduction of multi-user MIMO (MU-MIMO) systems in fourth-generation wireless systems. Multi-user information theory can now be seen as an important step towards the development of network information theory in the past decade or so, extending these concepts beyond single-hop multi-user networks. Both network coding and WPNC occupy the field of network information theory, and many concepts from it underlie the work in this book.

WPNC itself was discovered independently by three research groups, who approached it from slightly different angles, resulting in distinct approaches that, however, are clearly based on the same principles. Zhang, Liew, and Lam [64], of the Chinese University of Hong Kong, were probably motivated by concepts from network coding. They introduced the application of WPNC to the two-way relay channel, which we will review in the next chapter but which is quite similar to the butterfly network we have already seen. They also generalized it to a multi-hop chain network.

Popovski and colleagues at the University of Aalborg introduced an analog version of WPNC at the same time [49], based on earlier work applying network coding to the two-way relay channel [33]. They subsequently extended this to a scheme they refer