Chapter 24

BEYOND BIOLOGY

Melanie E. Moses and Stephanie Forrest

SUMMARY

1 Many of the greatest challenges facing science and engineering concern the flow of information, energy, and materials through networks. Examples include the spread of disease in increasingly interconnected human populations, the impact of fossil-fueled transportation on global climate, and advances in computation and telecommunication. The model of West et al. (1997, WBE) provides a unifying framework for understanding fundamental constraints, improving design, and predicting behavior in all of these complex systems.

2 Just as the cardiovascular network supplies energy to cells in organisms, so networks of transistors in computer chips support the information revolution, and road networks are the conduits through which people and goods move and interact to create vibrant modern cities. WBE offers insights into how networks govern the dynamics of these human-constructed systems.

3 Applying WBE in these settings reveals important commonalities and differences between biological and human-designed systems. The differences have led to important extensions and refinements of network scaling theory to account for issues such as: decentralized networks where resources do not flow from a single source; systems that become more densely populated as they increase in size; and modeling more carefully how resources travel from the terminus of a network to the components they service – the so-called "last mile."

4 These model enhancements have allowed us to apply WBE to human ecology and engineered systems, and they may lead to wider application of the theory in biology.

24.1 INTRODUCTION

In 1997 West, Brown, and Enquist (WBE: West et al. 1997) demonstrated how the branching architecture of the cardiovascular network generates the canonical metabolic scaling relationship, $B \sim M^{3/4}$ where B is the metabolic rate of an organism and M is its mass. The 3/4 exponent results from networks evolved to simultaneously maximize energy and resources delivered to cells, minimize the cost of transporting those resources, and minimize the cost of constructing and maintaining the network itself. The paper has been influential, not just because it proposed a mechanism to explain $M^{3/4}$ scaling, but also because it demonstrated how much of biology (described in previous chapters of this book) can be explained by understanding the flow of energy and other resources through networks. In this chapter, we show that the WBE approach can be extended even further to explain how networks constrain the design and growth of human-constructed systems, and in turn, how the topology and dynamics of engineered networks broadly affects how those systems function.

Metabolic Ecology: A Scaling Approach, First Edition. Edited by Richard M. Sibly, James H. Brown, Astrid Kodric-Brown.

We take as examples two different human-constructed systems: cities and computers. Cities are central to modern human ecology, with more than half of the human population living in urban areas. By even greater majorities, people in cities dominate the consumption of energy and materials and the production of new ideas, research, and inventions (Bettencourt et al. 2007). Computers and information systems are also central to modern human ecology, increasingly dominating our time, interactions with each other, and ways of solving problems. How do metabolic networks in biology relate to the flow of energy, information, and materials in cities and computers? In this chapter we discuss how the topology and dynamics of these networks constrain the way that humans move and process information, energy, and materials.

Although the examples of road networks and computer chips illustrate how WBE provides a unifying framework for understanding scaling in human-engineered systems, certain properties of these systems differ from cardiovascular networks and other biological resource distribution networks described by the WBE (West et al. 1997) model. Thus, applying WBE to these systems requires extensions and refinements to the original theory.

1 *Distributed networks.* Computer and road networks are less centralized than the cardiovascular system – there is not necessarily a single central source, like the heart, for all flow through the network. The theory can be corrected to account for multiple sources and destinations of information or resources within a network.

2 *Density dependence.* To a first approximation, the size and density of cells do not change with organism size within a taxonomic group. However, the density of transistors on computer chips has increased exponentially over time from thousands to millions of transistors per square millimeter, a phenomenon described by Moore's Law. Similarly, the density of people and businesses in cities increases with larger population size such that cities with a larger population have a higher density as well as a larger spatial extent (Samaniego and Moses 2008).

3 *The last mile.* Many networks deliver resources, energy, or information to a local service unit, and from this terminus the resource is transported to its final destination by other means, for example, by diffusion in the case of cells, or by wireless signals in the case of the Internet. We refer to this terminal service unit as "the last mile" by analogy with telecommunication networks. Work by Banavar and colleagues (2010) revised WBE to show how quarter-powers arise not from the fractal structure of the cardiovascular network, but instead from scaling of velocity to match the characteristic length of the capillary, or more generally, the length of the last mile.

4 *Accommodating superlinear scaling.* WBE demonstrates that, in centralized networks, the volume of a network that could deliver resources to each cell in a large animal as fast as they are delivered to a small animal must increase superlinearly with the organism volume (Fig. 24.1). This would require that the volume of the cardiovascular network grows faster than the volume of the animal. WBE assumes that this does not happen, consistent with the observation that mammals from mice to elephants are all approximately 8% blood (Peters 1983). Further, it is logically impossible for superlinear scaling to hold over large ranges in size: if

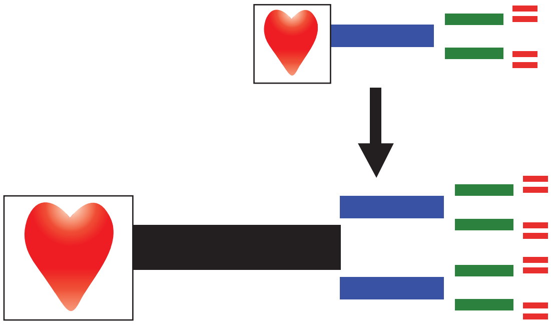

Figure 24.1 The top network shows a single aorta branching into two arteries that each branch into two capillaries. Following WBE each successive branch is shorter (in this case, 1/2 the length of the parent branch) and narrower (such that the summed cross-sectional areas of the daughter branches equals the area of the parent branch). In order to double the number of capillaries (from four in the upper network to eight in the lower network), the number of green arteries is doubled and the single blue aorta is duplicated. In order to obey the constraint that the blood flows through a single aorta, an *additional* branch must be added to the network. The black aorta in the larger network is both longer and wider than the other branches. Thus, the volume of red, green, and blue branches is doubled, but the volume of the larger network is more than doubled due to the addition of the black branch. In this way, the volume of the network grows faster than the number of capillaries that deliver resources: the network volume scales superlinearly with the network delivery rate.

the cardiovascular system in a mouse were 8% of the volume of the mouse, superlinear scaling would result in the cardiovascular system of an elephant being larger than the elephant itself. This is a key constraint in deriving the 3/4-power scaling exponent: because the network volume is constrained to scale linearly with the organism volume or mass, blood flows through the network at a sublinear rate. However, engineers have found several ways to accommodate superlinear network scaling in order to maintain fast delivery rates in large systems. These strategies, described below, accommodate superlinear network scaling over some range of system sizes.

To summarize, the scaling behavior of human-constructed systems poses a challenge for WBE because these systems have distributed networks, components whose densities vary with system size, different transport mechanisms to cross the last mile, and different strategies for overcoming the problem of superlinear network scaling. In the following, we give examples of how these challenges have been met in two quite different systems, cities and computer chips, and we then explore how these extensions of the theory may apply to biological systems, particularly ant colonies and other social animals.

24.2 CITIES

24.2.1 Scaling of road networks

Cities do not exist without roads. Just as the cardiovascular network distributes energy and materials to cells in an organism, urban roads are the arteries that transport people and goods to make the activities of businesses, households, and communities possible. Understanding the topology of urban networks that connect people and places provides insight into how cities are organized and may provide clues to how cities might be better designed to reduce traffic and increase interactions and innovation.

There are some similarities between urban road arteries and biological arteries. Cities have large highways and multi-lane surface streets that branch into successively smaller and slower boulevards that eventually deliver cars to surface streets, driveways, and parking lots. Roads should be designed so that surface streets that connect to highways have enough capacity to accommodate all of the exiting cars. When this ideal is not met, the result is a traffic jam. But traffic could be much worse than it actually is. If the topology of city roads matched that of the cardiovascular network, every city would have a central intersection downtown that every car passed through on every trip (Fig. 24.2A,B). In a sprawling city such as Los Angeles, each car would have to travel a substantial fraction of the radius of the city to drive to an enormous central intersection on every trip. Fortunately, the geometry of urban roadways and how drivers use them is not quite so simple. One usually does not require a trip across town to buy a gallon of milk or gasoline. In this sense, urban road networks are *less centralized* than biological vascular networks. There is not a single central place through which all traffic flows.

At the opposite extreme, we can imagine a completely decentralized city. In such a city, destinations are distributed evenly through the city, and no one would ever have to travel farther than the distance to their nearest grocery store, school, or restaurant (Fig. 24.2C,D). In such a city, the length of a trip would be determined not by the radius of the city (frequently tens of kilometers), but rather by the distance between destinations, perhaps a few blocks between coffee shops, or a few kilometers between shopping malls. In the latter case, per-capita transportation would depend entirely on density, and in dense cities like New York, people would hardly have to travel any distance at all.

Analysis of travel distances and road capacities (Samaniego and Moses 2008) indicates that cities are neither completely centralized (like the cardiovascular network and Fig. 24.2A,B) nor completely decentralized (as in Fig. 24.2C,D); they exist somewhere in between. Figure 24.3 shows that the average trip length is affected by both area and population density. Many people in large, densely populated cities travel very short distances to buy gasoline or groceries, and very long distances to commute downtown. Interestingly, in US cities, city area and density are correlated. That is, as cities (defined functionally as Metropolitan Statistical Areas) increase in population, they have both larger areas and more people and businesses per unit area. The greater area of large cities tends to increase travel distances, but this effect is somewhat mitigated by the increased density of large cities which reduces average driving distances. The scaling approach also points to a potential explanation for why traffic is more congested in large cities: empirically, per-capita driving distances are influenced by

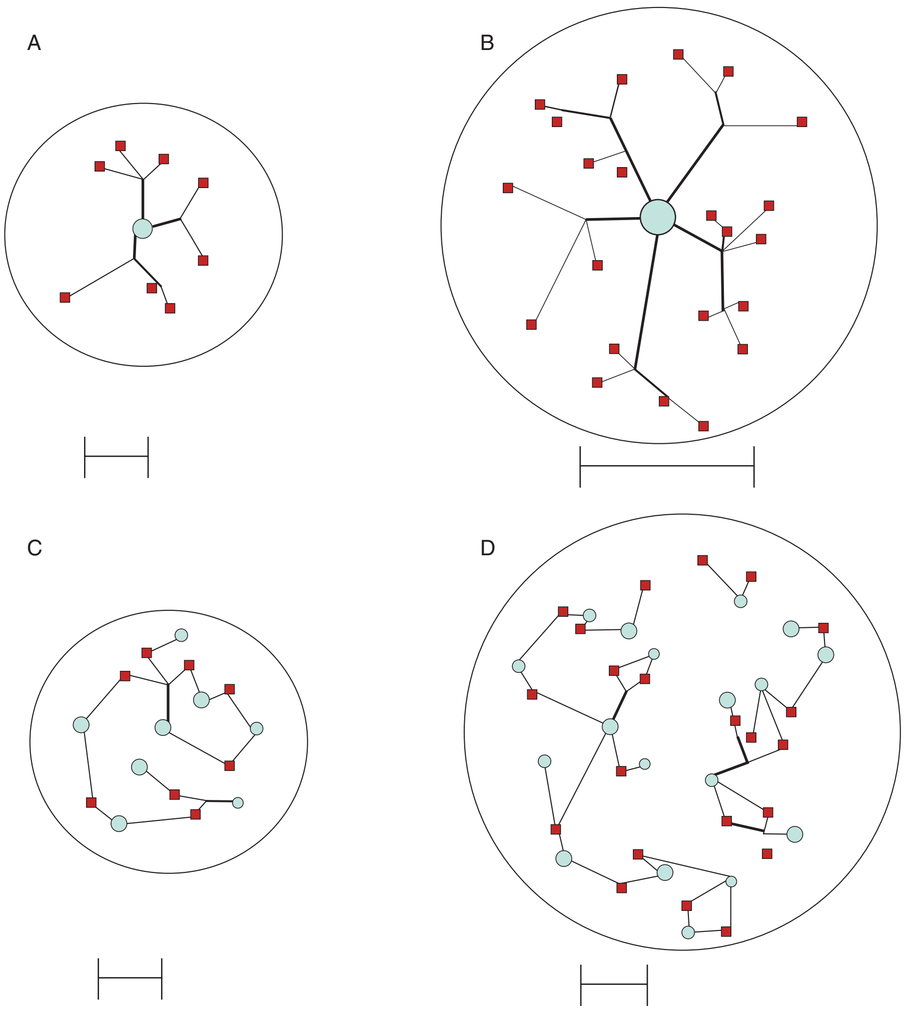

Figure 24.2 (A,B) In a completely centralized city every trip would originate at some location inside the city area (red squares) and end in a single downtown location (blue circles). The average trip distance would depend entirely on the area of the city. Since the city in (B) has twice the area of city A, average travel distances (shown below each figure) would be longer by a factor of the square root of 2. In a completely decentralized city (C,D), each trip would again originate from some location in the city area, but each trip would end at the nearest blue circle. Each blue circle represents a type of destination – a business or residence or park or other destination. In such a city, the area of the city has no impact on travel distances, only the density of destinations affects per-capita transport distances. Because city density increases with city size, the average distance traveled in city D is smaller than in city C. (Figure from Samaniego and Moses 2008.)

both city area and city density, but road capacity scales only with city density. The systematic deviation between road capacity and miles driven in large cities results in more congestion in larger cities. More generally, this approach offers a macroscopic perspective on the differences between small and large cities and on how road infrastructure and traffic might change as cities grow, and it gives urban planners a quantitative tool to understand the impact on traffic of different urban designs.

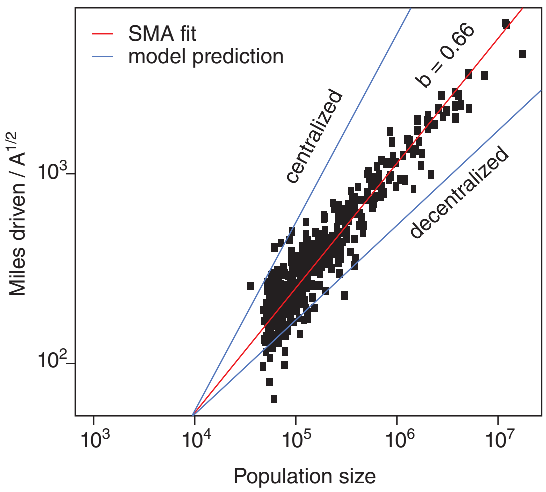

Figure 24.3 Predicted and observed relationships between total miles driven, population size, and area (A) of a city. The predicted relationship for centralized networks (as in Fig. 24.2A,B) is that the total number of miles driven is proportional to the population size multiplied by the radius of the city. This is equivalent to predicting that the number of miles driven divided by city radius ($A^{1/2}$) is linearly proportional to population size, as shown by the upper blue line with a slope equal to 1. The predicted relationship for decentralized networks (as in Fig. 24.2C,D) is that the total number of miles driven is proportional to population size divided by the square root of population density. Rearranging terms to show the decentralized prediction on the same axis as the centralized prediction, miles driven divided by city radius is predicted to be proportional to the square root of population size, shown by the lower blue line with a slope equal to 1/2. The empirical data from US urban areas shows that traffic follows an intermediate scaling that is influenced by both city area and population density.

24.2.2 Beyond infrastructure: scaling of human interactions in cities

Bettencourt and colleagues (2007) proposed that the size of a city systematically affects not just physical networks such as roads and power grids, but also the virtual networks of interactions between people. They consider how a wide range of city characteristics, such as rate of patent production, number of crimes, length of electrical cables, and total wealth of a city, relate to the size of the city population. They represent these relationships as scaling equations of the form $Y \sim N^{\beta}$ where Y is the city characteristic of interest, N is the city population size, and β is the scaling exponent. They find that wealth creation and innovation have superlinear scaling exponents, meaning exponents greater than 1, and often these exponents are near 1.15. Consequently, "larger cities are disproportionally the centers of innovation, wealth and crime, all to approximately the same degree" (Bettencourt et al. 2010). In contrast, scaling exponents related to physical infrastructure have sublinear exponents less than 1, and often close to the 3/4-power exponents seen in biology (Bettencourt et al. 2007).

These observations suggest that, as people are concentrated in larger cities, they interact more, and any phenomenon that occurs as a result of interaction happens more often for everyone. Both larger numbers and higher densities of people in larger cities may contribute to increased interactions. Each person in a large city has a greater chance to build on the ideas of others to generate a new idea, create a new business, and make more money. Since larger cities have more people and each person has more opportunity for innovation and wealth, there is a multiplicative effect, and innovation and wealth increase faster than population size. On the other hand, each person also has a greater chance to be involved in a crime or to spread a disease, and these phenomena are also disproportionally greater in larger cities. The theory allows predictions of how police forces, road maintenance costs, and other infrastructure and services should be expected to grow as population size grows in a city.

The Bettencourt et al. analysis provides a quantitative understanding of the human ecology of cities and is particularly relevant to understanding paths to sustainability. Because city size affects the rate at which people consume resources and generate creative new solutions and technologies, scaling in urban environments plays an important role in determining strategies for sustainable human populations (Moses 2009).

The authors also consider the effect of superlinear scaling exponents on urban growth. Metabolic growth equations for organisms show how sublinear exponents lead to asymptotic growth of animals to some maximum size (West et al. 2001; see also Brown and Sibly, Chapter 2). Systems with superlinear exponents could, in theory, grow indefinitely to infinite size. Given that cities and people living in them require resources that are limited, the theory predicts boom and bust cycles for cities, as are evident in empirical data. The theory also predicts that to avoid economic collapse, cities require increasingly faster cycles of innovation that increase the wealth and resource base of the city. In other words, as population increases, cities have to produce increasingly more wealth and innovation to avoid collapse.

This work also provides a way to quantify which cities are the most creative, the most violent, or most effective at generating wealth. Scaling exponents provide a baseline of expectation for a city given its size, so that individual characteristics of cities can be meaningfully compared. Just as allometric plots tell us that humans have unusually long lifespans for our size because they appear as outliers on an log-log plot, plots that take into account the expected increase of wealth and violence and innovation allow us to see which cities are particularly poor or innovative or free from crime. So for example Bridgeport, CT and San Francisco are particularly wealthy, and Corvallis, OR and San Jose, CA produce unusually large numbers of patents, given their size (Bettencourt et al. 2010).

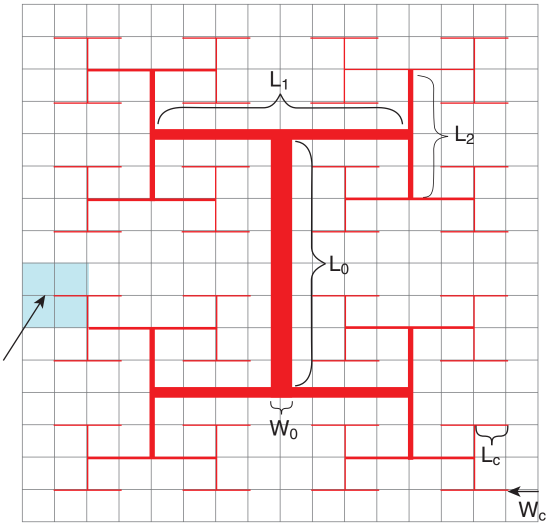

Figure 24.4 The H-tree is a fractal branching network that delivers a timing signal to all areas of a computer chip. Each successive branch is regularly shorter and narrower. The final branch delivers a timing signal to a service unit called an "isochronic region," shaded in blue.

24.3 COMPUTERS

24.3.1 Wire scaling on computer chips

Modern computer chips contain billions of transistors networked together in a few square centimeters of surface area. There are several distinct networks on these chips, which deliver power to the individual components, deliver a synchronizing clock signal, connect memory components and input and output, and implement the logic operations comprising the chip's functionality. As a concrete example, we will focus on the network known as the "clock tree," which delivers a timing signal to individual transistors, and we will show how clock trees and cardiovascular networks have similar topologies (Moses et al. 2008b). The clock tree is important because it consumes up to 40% of the chip's power, and it is a good example of the network scaling predictions from WBE once corrections are made to account for the clock tree being a two-dimensional rather than a three-dimensional structure. A common clock tree design known as the "H-tree" is shown in Figure 24.4. It is designed to deliver a timing signal from a clock to every location on the chip simultaneously. The topology of the H-tree precisely follows two branching rules proposed by WBE to minimize the resistance of flow through biological networks. When translated from three to two dimensions, those rules are:

1 *Wire width preserving.* When one wire splits into two daughter wires, their summed cross-sectional areas equal that of the parent.

2 *Area-filling branch lengths.* Each daughter wire is systematically shorter than its parent. In Fig. 24.4, the longest wire extends from the center of the chip, half way to the edge; the next wire goes from that point to half of the remaining distance to the edge, and so on. The length of each branch is the radius of the area to which it delivers a signal. The lengths are area-filling in that signal is delivered to all regions of the chip.

However, there is an additional assumption about cardiovascular networks in the WBE model which does not apply to H-trees. WBE assumed that capillaries have the same lengths across species, but the lengths of terminal wires of H-trees are different for large and small computer chips. This is because, empirically, the density of transistors on large chips is higher than the density on small chips: when transistors are closer together, the wires connecting them are shorter.[1] When this difference is accounted for, rules 1 and 2 precisely describe the branching geometry of the original H-tree design, shown in Figure 24.4, and are surprisingly similar to the depiction of the WBE model created by Etienne and colleagues (2006). These rules, along with other design constraints on computer chips, allow us to predict how much power is consumed by the H-tree and the whole computer chip. Additionally, a detailed analysis of the H-tree network sheds light on a debate about assumptions in the WBE model, leading to a more general model of biological network scaling (Banavar et al. 2010).

24.3.2 Scaling in information networks

The H-tree example shows that beneath some obvious differences (e.g., computer chips are measured in square millimeters, while animals are measured in grams or kilograms), there are striking similarities in the topology of distribution networks. There are also similarities in the function of these networks. Cardiovascular systems and H-trees are both infrastructure networks that connect components into a functioning system. Cells do not function without oxygen and nutrients, and transistors cannot compute without receiving electrons and timing signals from the clock tree. Just as the cardiovascular network dictates the pace of life, computational speed is constrained by H-trees and the logical "interconnect" that sends electronic ones and zeros to transistors.

Understanding this analogy requires understanding certain differences between computational networks and the resource distribution networks described by WBE. These differences include decentralized network flow, variation in the density of components, the role of the last mile, and accommodating superlinear network scaling. Once we account for the differences, we can use WBE network scaling principles to describe how power consumption, latency, and the physical footprint of a network scale as functions of system size, the number of components, and the degree of centralization.

Like the idealized cardiovascular network (Fig. 24.1), the size of the H-tree scales superlinearly: the footprint of the H-tree grows faster than the number of terminal wires that deliver timing signals to each isochronic region[2] (Fig. 24.4). However, there is an important difference in how the cardiovascular network and the H-tree accommodate superlinear network scaling. Because the original H-tree design (Fig. 24.4) required so much power to drive the long wide wires, engineers developed a way to minimize wire widths. Repeaters amplify signals so that even long wires have a small footprint. As the signal splits off at branch points, it is repeated, or amplified, to make up for losses. In that sense, there is no conservation of signal the way that there is conservation of blood. In this way, modern H-tree designs are modified to scale up more efficiently. It is noteworthy that amplifying a signal is possible in an information network, but amplifying energy (i.e., producing new blood cells or oxygen at intermediate points through the cardiovascular network) is not possible. Information can be amplified, but energy and materials cannot.

Engineers developed a second innovation to address superlinear wire scaling of the "interconnect," the network that forms electrical connections between logical elements on the chip. The footprint of the interconnect is simply allowed to consume a larger fraction of the surface of larger chips by adding metal layers of wire on top of the two-dimensional surface that holds the transistors (Moses et al. 2008b). Chips with more transistors have more metal layers to accommodate excess wires. Thus, while biological and computational networks face the same trade-off – the output of the network scales sublinearly with network size – the

[1] The exponential increase in transistor density has occurred because technological advances have made smaller transistors and thinner wires, allowing transistors to be packed more closely together. As the distances between transistors shrinks, wires connecting those transistors are shorter, so the density of components affects the size of the network connecting them.

[2] H-trees do not directly connect to every transistor. Rather, each terminal wire of an H-tree delivers a timing signal to an "isochronic region," the service unit that contains some number of transistors all receiving the same timing signal. When the isochronic regions are smaller and more dense, the frequency of the timing signal increases, and the chip can process data faster (this is the chip frequency that has increased from kHz to GHz and is often used to market computer chips). The density of this isochronic region affects the lengths of the H-tree network segments.

trade-off is managed differently. In biology, where superlinear network volume is not possible, the output is slowed in larger organisms. In two-dimensional computer chips, where economic pressures maximize speed (output), wire footprints that grow more quickly than chip surface areas are accommodated on additional surfaces.

These engineering innovations of moving wire onto metal layers and using repeaters to amplify signals have met market pressures that drive the design of faster and faster chips, with more and more transistors. However, those additional metal layers and repeaters consume power, and power has become a fundamental limit on modern chip design. By early in the twenty-first century, most PC chips consumed in the order of 100 watts (roughly the same metabolism as a human being) and more power was consumed in wires than in the transistors they connect. Miniaturization in transistor sizes has led to power-efficient transistors, but the wires that connect them scale in the opposite direction – their power consumption increases (Ho 2003).

Wire scaling became the fundamental problem in producing chips that conformed to Moore's Law – performing more computations with less power becomes impossible when wire scaling dominates power consumption. This problem is solved by decentralization. Just as decentralization allows cities to scale up more efficiently, exploiting decentralized networks improves scaling properties of computational networks. Our analysis suggests that much of the power consumed on chips can be predicted by assuming decentralized flow over the interconnect (Fig. 24.5) as indicated by the slope of the regression line that equals 1. This decentralization is being extended even further by the recent innovation of multi-core architectures; the process of placing increasing numbers of centralized processing units (CPUs) on a chip allows the interconnect to benefit from the same kind of locality that reduces traffic volume in decentralized cities (Fig. 24.2C,D).

Study of these decentralized networks suggests that wire scaling will remain an issue: as more cores are added to a chip, wire scaling *between* cores will dominate communication and power. There are ways to mitigate this problem by clever programming and architectures that maximize localized use of cores (Bezerra et al. 2010; Zarkesh-Ha et al. 2010). Just as people do not have to drive across town to buy a gallon of milk, electrons usually do not have to travel across an entire chip to compute.

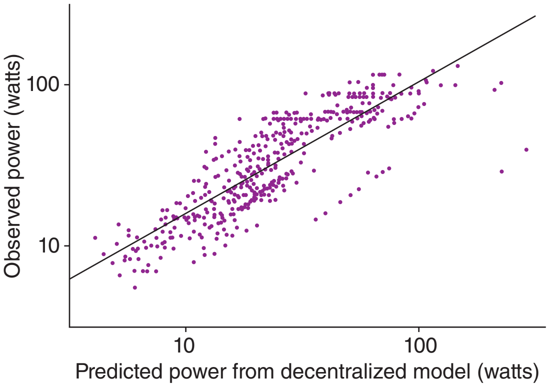

Figure 24.5 Measured versus predicted power consumed by microprocessors. Power is predicted assuming that wire length is determined by transistor density rather than chip area. Excluding the outliers on the right (diamonds representing chips with an unusually large number of transistors in the memory cache) the scaling exponent (slope of the regression line) is indistinguishable from 1 (Moses et al. 2008b).

The transition to multi-core chips mirrors the evolutionary transition to multicellular animals. The interconnect and H-tree on a single core becomes analogous to the transport of energy, materials, and signals within a cell, and the wires that connect cores become analogous to the cardiovascular network between cells. Interestingly, recent work in biology shows that the metabolic scaling exponent shifts in the transition from unicellular to multicellular organisms, the former having a nearly linear exponent and the latter having an exponent near 3/4 (deLong et al. 2010). Whether a similar shift in the scaling exponent will be seen in multi-core systems depends on whether engineers can continue to develop innovations to escape scaling constraints.

24.3.3 The last mile

Although the density of cells in animals is roughly constant, the density of transistors on computer chips varies by many orders of magnitude, from thousands to millions of transistors per mm^2. The enormous increase in transistor density over the last 40 years explains much (but not all) of "Moore's Law," an empirical observation that computing power has increased exponentially, doubling every 18–24 months over this period.

The absolute length of the isochronic region to which timing signals are delivered in computer chips is measured in nanometers, but despite its minuscule size, it determines the frequency of the clock and thus the information-processing power of the chip (Bakoglu 1990). Therefore, smaller isochronic regions and denser packing of both transistors and the terminal wire of the network lead to much faster speeds through the network. This is important because it means that the density of components alters network scaling.

The service unit has its analog in other transportation systems that distribute energy, materials, and information from a central source to dispersed locations and that have been designed to maximize performance. Examples in engineered networks are the last mile that connects individual consumers to global infrastructure networks such as the Internet or telephone networks or the electrical power grid. The service unit is also where a package flown across the globe at a speed of hundreds of miles per hour is walked to a door by a mail carrier, and where passengers exit high-speed planes and trains to take slower modes of transport home. Banavar et al. (2010) propose that the length of the service volume in an animal, proportional to the length of a capillary, also determines the speed of oxygen delivery and ultimately the scaling of metabolism. This model shows that the 3/4-power exponent is generated in networks without fractal structure because, in three dimensions, the radius of the organism scales as $M^{1/3}$, but the velocity is determined by the length of the last mile, which turns out to scale as $M^{1/12}$. This additional factor of $M^{1/12}$ transforms biological rates and times into "quarter"-powers of mass. The length and speed of transport over this last mile ultimately constrains delivery rates in a variety of biological and engineered networks.

24.4 BACK TO BIOLOGY

By applying network analysis and the MTE approach to human-constructed systems, we gain insights into how decentralization, density of components, and differences between networks of information and energy affect network scaling. These insights can be applied back to biology. Humans are not the only animals concentrated in dense populations with complex social structures and complex systems of information exchange. Recent work on ant societies shows that the size of an ant colony is related to its life history in the same way that an animal's body size determines its life history: a colony's metabolism, lifespan, and allocations to reproduction are all quarter-powers of colony size (Hou et al. 2010). This is particularly surprising because ant foraging networks are primarily two-dimensional, and so the scaling exponent might be expected to be lower (i.e., 2/3) if a two-dimensional network topology were governing colony metabolism in the same way that three-dimensional network topology governs organism metabolism. Evidence for "quarter"-power scaling in human socio-economic networks (Hamilton, Burger, and Walker, Chapter 20) suggests the same conundrum.

The surprising similarity in scaling patterns of organisms and "superorganisms" causes us to re-examine whether the topology of energetic networks is sufficient to explain 3/4-power scaling. The decentralized networks connecting ants in a colony are very much like roads in a city – they govern the rates of interaction among ants as well as the rates that ants collect resources and bring them back to a central nest. Larger ant colonies appear to have more interactions and communicate faster, and have more coordinated group foraging strategies, and more elaborate social structures than small colonies (Beckers et al. 1989). One hypothesis is that ants in larger colonies benefit from increasing returns in information exchange – that is, each ant in a larger colony is more informed and better able to share information about food locations, much as humans in cities are more able to exchange information to innovate and build wealth. The increased interactions and information exchange in large colonies might mitigate the diminishing returns of foraging networks that bring food to a central nest (Jun et al. 2003). However, a preliminary study shows that large and small colonies use information to improve foraging equally well (Flanagan et al. 2011). Why the networks in ant colonies that exchange information as well as transport energy and resources should lead to the same quarter-power scaling relationships seen in unitary organisms remains a mystery.

Understanding scaling properties of societies, including humans, ants, and other social organisms, requires a theory of how energy and information flow through networks. This has the potential to unify the study of organisms that are characterized as open energetic systems built upon networks that exchange both energy and information.