Mother Nature Knows Best: 
A Survey of Recent Results on Wireless Networks 
Based on Analogies with Physics 

S. Toumpis, University of Cyprus 

Abstract—We survey recent results on wireless networks that are based on analogies with various branches of Physics. We address, among others, the problems of optimally arranging the flow of traffic in wireless sensor networks, finding minimum cost routes, performing load balancing, optimizing and analyzing cooperative transmissions, calculating the capacity, finding routes that avoid bottlenecks, and developing distributed anycasting protocols. The results are based on establishing analogies between wireless networks and settings from various branches of Physics, such as Electrostatics, Optics, Percolation theory, Diffusion, and others. Many of the results we present hinge on the assumption that the network is massive, i.e., it consists of so many nodes that it can be described in terms of a novel macroscopic view. The macroscopic view is not as detailed as the standard microscopic one, but nevertheless contains enough details to permit a meaningful optimization.

Index Terms—Diffusion, Electrostatics, Geometrical Optics, Massive Networks, Massively Dense Networks, Multihop Networks, Percolation theory, Routing, Sensor Networks.

I. INTRODUCTION

In the design and analysis of communication systems and networks, researchers have often introduced analogies with Physics. Some examples are the application of fluid models in solving flow problems in networks [2], the use of free energy and the replica method of Statistical Physics in CDMA research [3], and, more famously, analogies between the entropy of Thermodynamics and the entropy of Information theory [4].

Until very recently, however, there had been no analogies proposed in the context of wireless networks. This is remarkable, since both physical phenomena and wireless networks have a prominent spatial aspect. For example, the placement of an electric charge affects the force it applies to all other charges around it. Similarly, in wireless networks the placement of a node affects the amount of power received by all other nodes in the network when the node transmits. In contrast, in wired networks nodes can move to arbitrary locations, provided the cables are long enough, without the network changing in any essential way.

In the last few years, however, there has been a surge of results in wireless networks proposing analogies with various disciplines of Physics. For example, it was shown that, under certain conditions on the physical layer, the optimal distribution of traffic in a wireless sensor network resembles a properly defined electrostatic field [5] (This result refined earlier but independent work in [6], [7]). As another example, as the number of nodes in a network goes to infinity, minimum cost routes converge to properly defined rays of light [8]. Examples of other areas of Physics that have been used are Particle flux theory (to calculate the traffic load), Diffusion theory (to devise efficient routing protocols), and Percolation theory (to study the capacity and connectivity of networks).

The aim of this survey is to introduce these research efforts to a wider audience, and so stimulate further research. We strive to present works that in many cases were arrived at independently by different researchers under a unifying point of view, but at the same time highlight their critical differences. We do not elaborate in any great length on the mechanics and details of the results, focusing instead on the critical ideas.

A common feature of many of the works we survey here is the assumption that the number of nodes is very large, technically approaching infinity. This assumption is nothing new; it was notably used in [9], with great success, to calculate the capacity of wireless networks in a certain setting. However, the works we survey, starting from [8], [7], [6] take this assumption to a conceptually higher level.

In particular, we assume that the number of nodes in the network is so large that, in addition to the standard microscopic view, an additional macroscopic view starts to emerge. The macroscopic view contains all the microscopic quantities that are needed for a complete description of the network: the locations of all nodes, the channel states between all pairs of nodes, the states of the buffers, and so on. As a result, performing network optimizations using the microscopic view is very complicated, due to the huge parameter space involved. The macroscopic view, on the other hand, is much more succinct, and is comprised of only a few, carefully selected macroscopic quantities, which describe the network with sufficient details to permit a meaningful optimization. The macroscopic quantities used depend on the amount of detail needed.

A typical macroscopic quantity, employed in many of the works we survey, is the node density function $d(r)$, measured in nodes per $m^2$. To define it, let us take a set $A$ of incremental size $|A|$, that contains the point $r = (x, y)$, and covering $N(A)$ nodes. The node density is defined as the limit

$$d(r) = \lim_{|A| \to 0} \frac{N(A)}{|A|}. \quad (1)$$

This work appeared, in preliminary conference format, in [1]. Contact information: Stavros Toumpis, Department of Electrical and Computer Engineering University of Cyprus, Kallipoleos 75, PO BOX 20537, 1678 Nicosia, Cyprus. email: toumpis@ucy.ac.cy. This work was funded by the EU IST Net-Refound project.
Formally, this limit is equal to 0 in all locations $r$, except in those where nodes are placed, where it is $\infty$. However, we assume that there is a range of sizes $|A|$ such that $A$ is (i) small enough to contain a uniform part of the network, but also (ii) sufficiently large so that the number of nodes within the set is very large and the ratio $N(A)/|A|$ is stabilized to a finite and positive value, which we take to be the limit.

We are clearly resorting to a hand-waving, however we stress that this is exactly the same hand-waving used in Physics to define, for example, charge density. Indeed, charge, whether positive or negative, is quantized. Therefore, if we consider a charged region and within it we consider a set of decreasing size, after a while the set will be so small as to contain no charges at all, and so the charge density at its center must formally be zero (unless the set is centered on a charge). Physicists routinely ignore this formality, and so will we from now on.

There is some disagreement regarding how to call networks where this novel macroscopic view is adopted: in [8] and subsequent works, they are referred to as massively dense. In [10], [11] they are referred to simply as dense, although this term has been used in different contexts. In [7] and subsequent works, no special name is given, and in [12] they are simply called very large. Calling these networks dense is perhaps misleading, as it suggests that each node has a very large number of neighbors within its communication range, which is not necessarily the case. On the other hand, some clear indication is needed that we are discussing about networks with a very large number of nodes. For these two reasons, in this work we refer to them as massive wireless networks.

It should be stressed that the macroscopic approach, although independently introduced in the context of wireless networks in [8], [7], [6], originally appeared in the community of road traffic engineers, as early as in 1952 [13], [14], and is still the subject of intense research within that community [15]. However, it is regrettable that the connection was not made until very recently, in [16], and until then the research activities of the wireless networking community had been totally independent of the previous results of road traffic engineers.

In Sections II, III, IV, and V we show how the massive network assumption has been used to address various problems in the design and analysis of wireless networks. In more detail, in Section II we present work on the optimization of traffic flow in wireless sensor networks, where, under specific conditions, the optimal flow resembles an electrostatic field. In Section III we examine an analogy between minimum cost routing and Geometrical Optics. In Section IV we review a formulation used in the analysis of cooperative transmission schemes in massive wireless networks. In Section V the problem of load balancing is defined, and useful bounds and expressions are presented, inspired from analogies with Particle flux theory.

In Sections VI, VII, and VIII we review various results that are based on analogies with Physics, but do not use the massive network assumption. In more detail, in Section VI we review a number of works that propose routing by use of electrostatic potentials. In Section VII a number of routing protocols inspired by diffusion processes are presented. Finally, in Section VIII we review results on the connectivity and capacity of wireless networks that are derived using Percolation theory.

II. Traffic Engineering and Performance Evaluation in Sensor Networks

In wireless sensor networks nodes are equipped with transceivers, used for communication over a common wireless channel, and, in addition, sensors, for collecting information about their surrounding environment, which acts as a distributed data source [17]. The sensed information must be delivered to a data sink that may or may not be distributed. Typically, nodes are immobile, small in size, and with very limited energy and processing resources. Applications range from the remote monitoring of plantations to intruder detection.

Within this framework, an important issue is designing the routes that data will take as they travel from the sources to the sink. On the one hand, we would like data to take relatively short routes, to minimize the use of resources. On the other hand, we also want to keep the levels of congestion down. Frequently, these goals are competing, leading to the formation of very interesting tradeoffs. In this section, we present a number of works that aim to optimize the shape of the traffic in massive wireless sensor networks.

A. Transporting traffic with the minimum number of nodes

The first problem we consider was proposed in [5]. We have available a set of wireless nodes that must be deployed in the region in order to support the transport of the data from the sources to the sinks. We assume that there is only one type of data, therefore any packet created at any of the sources can be delivered in any of the sinks. We assume that nodes communicate over a common channel of limited bandwidth, a reception is successful provided the Signal to Interference and Noise Ratio (SINR) is above a given threshold, and all collisions are averted by the use of an efficient media access protocol.
Our aim is to determine the optimal placement of the wireless nodes, so that the minimum number of nodes can support the delivery of all the created traffic. Also, we want to determine the flow of data traffic induced by this optimal placement.

**Macroscopic Quantities:** The first step is to specify our macroscopic quantities. The first one is the node density $d(x,y)$, measured in nodes per m$^2$, and defined by the limit (1). Therefore, in a location where $d(x,y) > 0$, there is a concentration of nodes such that the number of nodes in any set of incremental area $\epsilon$ centered at $(x,y)$ is $d(x,y)\epsilon$. The total number of nodes in the network, which we are asked to minimize, is given by the two-dimensional integral $N = \int d(x,y) dS$.

The next macroscopic quantity is the information density function $\rho(x,y)$, which is measured in bps/m$^2$ and jointly describes the distributed sets of sources and sinks. In a location $(x,y)$ where $\rho(x,y) > 0$, there is a distributed set of sources, which jointly create traffic with rate $\rho(x,y)\epsilon$ over any set of incremental area $\epsilon$ centered at $(x,y)$. Similarly, in a location $(x,y)$ where $\rho(x,y) < 0$, there is a distributed set of sinks, which jointly require traffic with rate $\rho(x,y)\epsilon$ over any set of incremental area $\epsilon$ centered at $(x,y)$.

The last macroscopic quantity needed is the traffic flow function $T(x,y) = T_x(x,y)\hat{x} + T_y(x,y)\hat{y}$, which is a vector function measured in bps/m and models the flow of traffic at location $(x,y)$. Its direction gives the direction of the traffic flow, and its magnitude gives the intensity of the traffic. In more detail, consider a line segment located at $(x,y)$, oriented so that it is vertical to the traffic flow, and of incremental length $\epsilon$. The magnitude $|T(x,y)|$ is such that the traffic flowing across the line interval equals $|T(x,y)|\epsilon$.

When viewing a specific location of the network, one may observe many distinct streams of traffic, possibly along different directions. However, the fact that the traffic streams all carry the same type of packets allows us to combine them by performing vector addition, and thus abstract the movement of data at the microscopic level by a simple macroscopic quantity, the traffic flow function at that location. To clarify things, suppose as an example that at a particular location of the network there are two traffic streams carrying equal amounts of data packets in opposite directions. Although in the microscopic level the network the network is transporting data, on the macroscopic level the two streams cancel out, and the network appears inactive. Indeed, the nodes of that location do not contribute to the net accumulation of data packets at the destination.

**Constitutive Laws:** Having defined our macroscopic parameters, we specify two constitutive laws that must connect them. To justify the first law, note that the traffic flow function $T(x,y)$ describes the way the traffic moves across the network. The information density function $\rho(x,y)$, on the other hand, describes how the traffic is created or absorbed. Clearly, the two must be related, to ensure that together they present a consistent picture of the flow of packets. For example, if the traffic flow shows data flowing out of a given location $(x,y)$, the information density at that location must have the positive value that exactly matches the rate of the outflow.

Mathematically speaking, this requirement translates to

$$\nabla \cdot T = \rho,$$

where $\nabla \cdot T \triangleq \frac{\partial T_x}{\partial x} + \frac{\partial T_y}{\partial y}$ is the divergence$^1$ of $T$.

To derive the second constitutive law, the first step is to observe that the node density $d(x,y)$ needed to support the traffic flow is:

$$d(x,y) = k|T(x,y)|^2,$$

where $k$ is a constant that depends on the physical layer. This relation is intuitive: if there is no traffic flowing through a region, then we do not need to place any nodes there. Also, the more traffic flows through a region, the more nodes are needed. But why choose this particular relation and not, say, $d(x,y) = k\langle T(x,y) \rangle^2$? In [5], it is shown that (2) is well justified for the particular physical layer adopted. Using (2), the minimization problem now becomes

$$\minimize: k \int |T(x,y)|^2 dS,$$

from which it can be shown in a straightforward manner (the proof can be found in [6], [7], [5]) that

$$\nabla \times T = 0,$$

where $\nabla \times T \triangleq \frac{\partial T_y}{\partial x} - \frac{\partial T_x}{\partial y}$ denotes the rotation of $T$. (Note that in the two-dimensional space the rotation of a vector function is scalar.)

**Analogy with Electrostatics:** Up to now, we have defined the functions $T(x,y)$, and $\rho(x,y)$, and we have shown that:

$$\nabla \cdot T = \rho, \quad \nabla \times T = 0.$$

We now make the critical observation that this pair of equations is identical, up to a multiplicative constant, to the set of equations that describes the electrostatic field $E(x,y)$ created in two-dimensional free space by a distribution of charge $\rho(x,y)$:

$$\nabla \cdot E = \frac{\rho}{\epsilon_0}, \quad \nabla \times E = 0,$$

where $\epsilon_0$ is the permittivity of free space. Therefore, the optimal traffic flow, that minimizes the number of nodes needed, happens to be the same as the electrostatic field that will be created if we remove the data sources and substitute them with positive charges, and we remove the data sinks and substitute them with negative charges.

As an example, in Fig. 2 we plot the electrostatic field created by three distributed positive charges and a single distribution of negative charge. The distributions are such that the net charge is zero. By the analogy, the electric field lines coincide with the optimal routes that traffic must take to go from three distributed data sources to a single distributed data sink, in order to use the minimum number of nodes. As expected, most of the traffic will follow relatively short routes, but some of the data will follow long routes, in order to avoid excessive congestion at the central part of the network.

---

$^1$For a detailed derivation of this equation, we refer the interested reader to [5]. For a detailed introduction to divergence and related operators, we refer the reader to [18].
The analogy is very tight, and can be extended toward many different directions, which are explored in detail in [5], [19]. For example, the electrostatic potential difference between two locations A and B, translated in our networking setting, becomes the number of hops between A and B. Also, the electrostatic field energy of an electrostatic setting, translated in the networking setting, becomes the total number of nodes. Electrostatic setting with different dielectric materials correspond to nonhomogeneous networks where various regions have varying data transporting capabilities, so that the constant $k$ of (2) is a function of the location. More impressively, by using Thomson’s law [20] we can show that electrostatics settings with conductors correspond to networks with sources and sinks that have limited mobility, and in particular can freely move within the conductors. In order to minimize the number of nodes needed, the sources and sinks will position themselves in the conductors in exactly the same way that charges would position themselves in order to minimize the electrostatic field energy. Finally, sensor networks with boundaries through which traffic can not flow correspond to electrostatic settings where a Neumann boundary condition is taken on the boundary. (This is the case with the boundary of Fig. 2.)

\[ \nabla \cdot \mathbf{T} = \rho. \] (4)

\[ \text{minimize: } \int |\mathbf{T}(x,y)|^2 dS. \] (5)

After a few manipulations, this requirement is shown to be equivalent to the condition

\[ \nabla \times \mathbf{T} = 0. \] (6)

As in Section II-A, the combined set of equations (4) and (6) shows that the traffic flow is identical to the electrostatic field that would be created if we removed the data sources and sinks, and substituted them with positive and negative electrical charges.

The minimization problem (5) implicitly assumes that all regions of the network must be equally protected from unduly heavy traffic. If, however, some regions of space are more sensitive to high levels of traffic that other regions, for example because they have fewer energy reserves, it is more appropriate to require:

\[ \text{minimize: } \int K(x,y)|\mathbf{T}(x,y)|^2 dS. \] (7)

$K(x,y)$ is a scalar function which is given high values in the regions of the network that must be protected from high levels of traffic. After straightforward algebra, it is shown that this minimization is achieved for the traffic flow that satisfies the condition

\[ \nabla \times (K \mathbf{T}) = 0. \] (8)

The traffic flow that satisfies both (4) and (8) again resembles an electrostatic field, and in particular the electric displacement field that exists within a nonhomogeneous dielectric material with relative permittivity equal to $[K(x,y)]^{-1}$.

The authors evaluate by simulation the performance of routing protocols where the routes are derived from these analogies with Electrostatics and, for the particular network models and topologies that are simulated, important gains in energy efficiency and network lifetime are reported. However, and although the minimization problems (5) and (7) are very intuitive, they are not formally shown in [7] to be the correct problems to solve. In fact, as was shown later on in [19], for some physical layers they provably are appropriate, but for other physical layers they provably are inappropriate.

The formulation until now crucially depends on the assumption that there is a single type of traffic, and therefore a single type of sources and sinks. If this is not the case, then we need multiple functions $\mathbf{T}^i(x,y)$, $i = 1, 2, \ldots, M$, and $\rho^i(x,y)$, $i = 1, 2, \ldots, M$. In [21], the authors propose the following optimization problem

\[ \text{minimize: } \int K(x,y)\mathbf{T}^T \mathbf{H} \mathbf{T} dS, \]

subject to:

\[ \nabla \cdot \mathbf{T}^i = \rho^i, \quad i = 1, 2, \ldots, M, \]
where the tensor $T = [T^1 | T^2 | \ldots | T^M]^T$ is a column of vectors of length $M$, and $H$ is an $M \times M$ positive definite matrix that captures the effects of interacting traffic streams on the consumption of energy. Note that this optimization problem does not capture all cases of interest; in particular, there is no matrix $H$ such that $T^T H T = (|T^1| + \ldots + |T^M|)^2$. Nevertheless, it is a highly non-trivial generalization of the single commodity case. It would seem at first that this optimization problem is insurmountable. However, by a clever application of Singular Value Decomposition it can break down into $M$ independent minimizations of integrals with the form of (7).

An alternative, very promising derivation for handling multiple classes of traffic has recently appeared in [16]. This work is inspired by previous work on road traffic networks and makes the assumption that the use of directional antennas in the network constraints all transmissions to be either in the North-South or East-West direction. The authors derive conditions for a traffic flow consisting of $\nu$ classes to be optimal.

Finally, we note in [22] the authors consider the case where there is a single type of traffic, which must be delivered to not one but a set of sinks and, in addition, the amount of traffic that must arrive in each sink is also subject to optimization. The authors show that in order to solve the problems (5) and (7), each sink should receive so much traffic that, in the equivalent Electrostatics setting, all sinks have the same electrostatic potential.

C. General Case

Motivated by the optimization problems of the previous sections, we now consider the following problem, which first appeared in [19]:

\[
\begin{align*}
\text{minimize:} & \quad \int_A G(x, y, |T(x, y)|^2) \ dS, \\
\text{subject to:} & \quad \nabla \cdot T(x, y) = \rho(x, y). 
\end{align*}
\]  

(9)

Therefore, we associate with a volume of traffic $|T(x, y)|$ traveling across location $(x, y)$ a cost density $G(x, y, |T(x, y)|^2)$. We would like to determine the optimal traffic density $T(x, y)$ that minimizes the cost across the whole network, while at the same time ensuring that indeed traffic is created and disappears only at prescribed places, and with prescribed rates, i.e., $\nabla \cdot T(x, y) = \rho(x, y)$.

We note that the problems (3), (5), and (7) are special cases of (9). Also, since the cost density depends explicitly on the location $(x, y)$, it can also capture the cost of sensing, and also non-homogeneous networks. On the other hand, there is nothing in the way the problem is formulated to capture the effects of anisotropic networks, i.e., networks that behave differently along different directions. (Such is the network in [16].) In any case, the problem is sufficiently abstract to be applicable in a wide variety of transportation settings, also outside the context of wireless networks.

Using standard Calculus of Variations tools, it is shown in [19] that the optimal traffic flow that solves (9) is given by the equation

\[
T(x, y) = \frac{1}{2G'(x, y, H(x, y, |\nabla \phi|))} \nabla \phi,
\]

where the scalar potential function $\phi(x, y)$ satisfies the following nonlinear partial differential equation:

\[
\nabla \cdot \left( \frac{\nabla \phi}{2G'(x, y, H(x, y, |\nabla \phi|))} \right) = \rho.
\]

The functions $G'$ and $H$ are derived from $G(x, y, |T(x, y)|^2)$ in a straightforward manner.

As a numerical example, in Fig. 3 we plot the optimal traffic flow for a network in which three distributed data sources send data to a single distributed data sink. The network has an orthogonal boundary that data packets cannot escape; this imposes an additional constraint that appears as a boundary condition on the potential function. We assume that $G(x, y, |T(x, y)|^2) = |T(x, y)|^{2a}$, and we plot the cases $a = \frac{3}{4}, \frac{4}{5}$. The choice $a = 1$ corresponds to the Electrostatics case and appears in Fig. 2. When $a = \frac{3}{4} > 1$, the cost increases with the traffic faster than in that case, and so areas with high traffic flow are penalized more. Therefore, the optimal routes are longer than in the Electrostatics case. The opposite happens...
in the case $a = \frac{3}{2} < 1$: the network can handle traffic better that in the Electrostatics case, and as a result the routes become shorter and more direct.

We briefly note that an optimization problem similar to (9) recently appeared in [16]. There, the authors adopt a game theoretic approach, assuming that each packet is an infinitesimally small player, attempting to minimize his own transportation cost. Naturally, the cost of each player depends on other player’s decisions. In such a setting, the authors show that a Wardrop equilibrium exists, so that all used routes between a source and a destination have the same minimum cost, and all unused routes have costs equal or greater than that minimum.

D. Performance Evaluation of Wireless Sensor Networks

In the works presented so far in this section the specific aim was to minimize the total cost in the network, given in terms of a two-dimensional integral, by optimizing the flow of traffic. In contrast, the authors of [23], [24] consider a more open-ended task, namely the evaluation of the performance of a wireless sensor network as various parameters change.

The overall methodology is as follows: First, the authors develop a very detailed microscopic model of the network. In particular, they specify a precise model for the energy needed for communication, as well as the conditions for a transmission to be successful. All traffic must be delivered, in a multihop fashion, to a centrally located sink, and with the minimum amount of energy. Nodes use Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) for performing media access, and are allowed to enter a low power sleep state.

Secondly, the authors define a set of macroscopic quantities, to be used for evaluating the performance of the network. Among others, the authors define the following:

- The sensor density $\rho(r)$, measured in sensors/m$^2$.
- The traffic generation density $\lambda(r)$, measured in bps/m$^2$.
- The total traffic density $\Lambda'(r)$, measured in bps/m$^2$, which includes the routed traffic.
- The actual total traffic density $\Lambda^*(r)$, which includes the effects of retransmissions.
- The probability density $u(r'|r)$ which equals the probability that packets created at $r$ will be routed through $r'$.
- The mean delivery delay $D(r)$ for packets starting at $r$.

Thirdly, the authors use the detailed microscopic model to derive closed form expressions involving macroscopic quantities, such as the actual total traffic density and the mean delivery delay, with microscopic parameters, such as the communication range, the percentage of time that nodes remain in the sleep state, etc.

Finally, the authors use these expressions to evaluate the performance of the network (given in terms of macroscopic quantities), as a function of the various microscopic parameters.

With respect to previous works, this work is much more ambitious, as it involves a more detailed microscopic network model, and a larger number of macroscopic parameters. As a result, some results are numerical and not in closed form.

On the other hand, the macroscopic parameters capture a lot of information, giving a much more complete picture of the network.

III. Minimum Cost Routing in Wireless Ad Hoc Networks

In this section we present an analogy between optimal routing in wireless networks and Geometrical Optics. The analogy, which critically hinges on the massive network assumption, first appeared in [8], and was later extended in [25], [26], [27], [28].

To motivate readers, let us consider the example of Fig. 4, where we plot the minimum-cost route connecting a source node placed at $(-50 \text{ m}, -37 \text{ m})$ and a destination node placed at $(50 \text{ m}, 37 \text{ m})$, through an area where relay nodes are placed according to a spatial Poisson process of density $\lambda(x, y) = a \times [0.1 + \exp(-0.1|x|)]$ nodes per m$^2$, for four increasing values of $a$ ($a = 0.1, 0.3, 1, 3$). The cost connecting two nodes that are separated by a distance $l$ equals $c(l) = l^2$. Obviously, in order to calculate the minimum-cost route, we need the precise location of all nodes in the network. As the figures reveal, however, as the number of nodes increases, the optimal route starts more and more to resemble a continuous curve, and it is intuitively clear that the shape of the curve does not depend on the precise node placement, but only on the node density function $\lambda(x, y)$ and the cost-versus-distance function $c(l)$. We now formalize this observation, using an analogy with Geometrical Optics.

A. Analogy with Geometrical Optics

To derive the analogy, we model the network in terms of a single macroscopic quantity, the cost function $c(r)$, defined as the following limit:

$$c(r) \triangleq \lim_{\epsilon \rightarrow 0} \frac{dc(\epsilon, r)}{\epsilon \times c_{\text{nominal}}},$$

where $dc(\epsilon, r)$ is the cost of transporting a packet over an incremental distance $\epsilon$ at the neighborhood of the location $r = (x, y)$ and $c_{\text{nominal}}$ is the cost of transporting a packet over a uniform, nominal network.

In the massive network limit, routes resemble continuous lines, and consequently the cost of transporting a packet from point $A$ to point $B$, over a non-incremental route $\mathcal{R}$, is given by the line integral

$$[A, B]_{\mathcal{R}} \triangleq \int_{A}^{B} c(r) \, dr$$

over the curve $\mathcal{R}$. Introducing the concept of the nominal network and dividing with its cost $c_{\text{nominal}}$ may seem unnecessary. However, there are two important advantages: firstly, in this manner the cost function is unit-less, irrespective of the precise nature of the cost. This makes the definition more general. Secondly, we will soon show an analogy with Optics, where the cost function is mapped to the refractive index. As the refractive index is unit-less, making the cost function also unit-less makes the analogy tighter.
Unicast Routing: The first routing problem we pose is the determination of the optimal route $R_{opt}$ connecting points $A$ and $B$ that incurs the minimum transport cost, as this is given by (10). Mathematically speaking, our problem involves the minimization of an integral along a curve, over all curves that satisfy certain constraints (i.e., the curves must start in $A$, end at $B$, and be continuous). In principle, such questions can be tackled by applying Calculus of Variations. However, this problem is actually solved by nature, and so has already been studied extensively.

In particular, let the cost function $c(r)$ be the refractive index of an optical medium [29]. Accordingly, the cumulative route cost $[A,B]_{R}$ becomes the optical length [29] of the curve $R$. Now let a source of light exist at point $A$. The principle of Fermat states that the rays of light connecting $A$ with any other point $B$ are exactly those curves that have the property of having optical lengths that are local minima (i.e., if we consider a small perturbation of a ray of light, its optical length will necessarily be larger than the optical length of the ray). In general, there can be many rays of light connecting the two points. Each of them represents a local minimum, and therefore one or more of them will also be a global minimum.

Moving back to the networking context, it follows that a curve that connects two locations $A$ and $B$ is a local minimum if and only if it coincides with a ray of light connecting the two locations, if we replace the network with an optical medium of refractive index $c(r)$. Therefore, we only need to look among all rays to find the global minimum. As the propagation of rays of light is very well studied and understood, we can use the knowledge and intuition already accumulated for our own problem. For example, we know that rays of light tend to bend toward optically denser regions of space (i.e., regions with high refractive index), so as to satisfy the following differential

![Fig. 4. Minimum cost routes in a network with spatial node density $\lambda(x,y) = a \times [0.1 + \exp(-0.1|x|)]$ nodes per m$^2$. (a): $a = 0.1$. (b): $a = 0.3$. (c): $a = 1$. (d): $a = 3$.](image)
Fig. 5. (a) The three rays that arrive at the location $B = (10 \, m, 30 \, m)$ if a source is placed at the location $A = (0 \, m, 0 \, m)$. The refractive index is
\[ n(x, y) = 1 + \exp[-0.1 \sqrt{x^2 + (y - 10)^2}] \]
The optical lengths of the routes are locally minimum, but only the optical length of the ray passing on the right of the region with higher refractive index is a global minimum. Areas with high refractive index appear darker. (b) Light rays and loci of the eikonal $S(r) = 100 \times i$, $i = 1, 2, \ldots$ for a medium with refractive index $n(x, y) = [(3 \times 10^{-5})x^2 + 0.02^{-2}]$ and a circular source of radius 10 m centered at $(50 \, m, 100 \, m)$. Areas with higher refractive index appear darker. (c) Three optimal routes connecting the locations $(-50 \, m,-37 \, m)$ and $(50 \, m, 37 \, m)$, in a network with node density $\lambda(x, y) = [0.1 + \exp(-0.1 \times |x|)]$ nodes per $m^2$. Each route is optimal under a different cost function. Areas with higher node density appear darker.

equation:
\[ \frac{d}{ds}(n \, dr) = \nabla n, \quad (11) \]
where $r = (x, y)$ is the position vector of a typical point on the ray, $s$ is the length of the ray measured from a fixed point on it, $dr/ds$ is the direction on the ray at $r$, $n(r)$ is the refractive index at $r$, and $\nabla n \equiv \frac{\partial n}{\partial x} + \frac{\partial n}{\partial y}$ is its gradient. The very same equation must be satisfied by the minimum-cost routes, if we substitute the refractive index $n(r)$ with the cost function $c(r)$.

As an example, in Fig. 5(a) we plot the three rays that arrive at the location $B = (10 \, m, 30 \, m)$ if a source is placed at the location $A = (0 \, m, 0 \, m)$. The rays are constructed by picking an initial angle of departure from the source and then performing raytracing using (11). (Note that the source is launching rays for all possible angles of departure, however only these three go through the point $B$.) The refractive index is $n(x, y) = 1 + \exp[-0.1 \sqrt{x^2 + (y - 10)^2}]$. Therefore, the medium has a region with high refractive index (i.e., a lens) centered at the location $(x, y) = (0 \, m, 10 \, m)$. All three associated optical lengths are local minima, however only one of them is the global minimum. In the context of networking, the figure shows the three locally optimal routes connecting two locations $A$ and $B$, of which one is also globally optimal.

**Anycast Routing:** We now consider a more general problem: let $\Omega \subset D$ be a sub-domain of $D$ containing a subset of nodes in the network. Given any other node $B$ in the network, what is the minimum cost route connecting $B$ with any of the nodes in $\Omega$, at which node in $\Omega$ does it end, and what is the associated minimum cost?

This problem is relevant when we need to find a minimum cost route, but we have some freedom in selecting one of its end points. For example, the set $\Omega$ might be the set of the locations of all cluster heads of a wireless network using clustering. In such a case, each node will want to communicate with the cluster head which leads to the minimum cost.

This problem can also be solved by invoking the Optics-Networking analogy. In particular, we need the minimum costs, and the routes that achieve them, between the region $\Omega$ and all other points in the network. In the context of Optics, we need to find the minimum optical lengths, and their corresponding rays, between a distributed light source occupying $\Omega$ and all other points, in an optical medium whose refractive index is $n(r) = c(r)$. By definition, these minimum optical lengths are given by the eikonal function $S(r)$, which is found by solving the following eikonal equation:
\[ (\frac{\partial S}{\partial x})^2 + (\frac{\partial S}{\partial y})^2 = n^2 \iff (\nabla S)^2 = n^2 \iff |\nabla S| = n. \quad (12) \]
In the above, $n(r)$ is the refractive index at $r$, and $\nabla S \equiv \frac{\partial S}{\partial x}\hat{x} + \frac{\partial S}{\partial y}\hat{y}$ is the gradient of $S(r)$. The boundary condition that accompanies the eikonal function is that $S(r) = 0$ on all points $r$ that belong to the boundary $\partial \Omega$ of the source set $\Omega$. It can be shown that the light rays emanating from the distributed source are constantly orthogonal to the loci of constant values of the eikonal function.

As an example, in Fig. 5(b) we plot the rays of light and constant loci $S(r) = 100 \times i$, $i = 1, 2, \ldots$ for a medium with refractive index $n(x, y) = [(3 \times 10^{-5})x^2 + 0.02^{-2}]$ and a circular source of radius 10 m centered at $(50 \, m, 100 \, m)$. In the context of networks, the light rays become optimal routes and the loci become contours of constant minimum cost.

We note that the eikonal equation is nonlinear, hence the eikonal function can have multiple values. In locations where this is the case, each value corresponds to the locally minimum optical length of a distinct light ray arriving at that location. As we are interested in globally minimum optical lengths,
whenever there are multiple values we only maintain the smallest, and its associated light ray.

**Broadcast Routing:** Let $\Omega \subset \mathcal{D}$ contain a subset of the nodes in the network. If these nodes hold a data packet that must be received by all other nodes in the network, what is the optimal manner with which the packet must propagate through the network, in order to minimize the total routing cost?

Similarly with the previous case, it is shown in [27] that a data packet initially existing in a subset $\Omega$ and optimally propagating over the whole network resembles light propagating from a distributed source $\Omega$, if the cost function is taken to be the refractive index of the medium. As an example, Fig. 5(b) also shows how a packet originating in a circular source is optimally broadcast through the whole network.

### B. Selecting a Cost Function

Until now, we have taken the cost function as a given, without considering how it can be reasonably defined. For this, a detailed specification of the network is needed. Clearly, the model will specify the placement of the nodes in the network, and the cost $c(l)$ incurred by transmitting a packet to a neighbor at a distance $l$. More importantly, the model must contain a forwarding rule that nodes will use to select the next hop of a packet, assuming that the macroscopic curve that the packet will follow is specified. This forwarding rule is critical, and in fact various such rules have already been independently proposed under the name of Trajectory Based Forwarding (TBF) [30], [31], [32], [33], in networks that are not massive.

As an example of a microscopic model, in [27] it is assumed that nodes are placed according to a spatial Poisson process with node density $\lambda(r)$. Also, if a packet is stored in a node $A$, and must be transported along a curve $R$, then the forwarding rule specifies that the next node to receive the packet is the one for which the deviation from the macroscopic route is not increased and in addition the quantity $\sqrt{\lambda}$ is minimized, where $C$ is the cost of transmitting to the node, and $P$ is the resulting progress that the packet will make. The progress is measured in meters, and equals the projection of the vector connecting the transmitter and the receiver on the macroscopic route. Finally, it is assumed that the bandwidth in the network is limited, and so must be used judiciously. Following the work of [9], a reasonable model for the transmission cost, is $c(l) = l^2$. Using all these assumptions together with the massive network assumption, it is derived in [27] that the cost function is:

$$c(r) = \frac{1}{\sqrt{\lambda(r)}}. \quad (13)$$

Therefore, the cost is smaller in regions where the node density is large. The reason is that if there are many nodes, the average distance between transmissions, and the associated transmission cost, will be small.

The cost function (13) is only one of many possibilities. For example, in [27] the authors show that in networks where the available bandwidth is very large, but the energy available to nodes is limited, a more appropriate cost function is

$$c(r) = f(\lambda(r)), \quad (14)$$

where the function $f(x) \to \infty$ as $x \to 0$, and $f(x) \to \text{const} > 0$ as $x \to \infty$. As another example, in [8], where the analogy with Optics was first considered, the author assumes a setting in which the number of hops must be minimized, and in addition communication is constrained between nearest neighbors. In this setting, the author implicitly uses the cost function

$$c(r) = \sqrt{\lambda(r)}. \quad (15)$$

As a final example, in [34] the authors propose, independently from this line of work, a cost function that captures the energy needed for routing in a large-scale static interference field. Clearly, many other microscopic network models can be considered, each of them leading to a different cost function.

As is intuitively clear, the choice of cost function can drastically affect the shape of the optimal routes. As an example, in Fig. 5(c) we plot the optimal routes connecting the locations $(-50 \text{ m}, -37 \text{ m})$ and $(50 \text{ m}, 37 \text{ m})$, in a network with node density $\lambda(x,y) = [0.1 + \exp(-0.1 \times |x|)] \text{ nodes per m}^2$ and the cost function is given by (13) (resulting in route R1), (15) (resulting in route R2), and by the equation $c(r) = \text{const}$ (resulting in the straight route R2).

### C. Open Research Issues

Despite its usefulness in characterizing the optimal routes, the Optics analogy does not provide any method that the nodes can use to determine these routes in a practical, distributed manner. Mathematically speaking, although the nodes know that the optimal routes satisfy the differential equation (11) and the eikonal equation (12), they lack the initial angles with which the routes should be launched from the sources and, in the cases of the anycasting and broadcasting problems, also the initial launching point.
In the context of unicasting, a simple solution to this problem, proposed in [26], is the following: Each node that is interested in participating in an exchange of packets, launches Route Request (RREQ) packets along a number \( N \) of exploratory rays separated by equal angles, specifying its identity, and the identity of all the nodes with which it would like to exchange packets. Nodes that receive RREQ packets from a pair of nodes interested in exchanging packets will notify them about each other’s location, using the rays that have already been propagated. An example appears in Fig. 6, for a network where the cost function is given by \( c(x, y) = 1 + 2 \exp[-y^2] \). Similar ideas have appeared in the context of location discovery protocols used in conjunction with geographical routing algorithms [35].

This discovery scheme has a number of important advantages: Firstly, nodes need not know the physical location of the nodes with which they want to communicate. Secondly, the shape of the exploratory routes does not need to be determined at the source. Rather, the nodes of a location through which the exploratory routes passes will bend it incrementally, according to (11) and the local properties of the medium. Finally, the scheme can be extended in a straightforward manner to also cover the cases of anycasting and broadcasting.

On the other hand, the routes that the nodes find are not optimal, but rather consist of two optimal segments. In addition, a potential complication is that rays that emanate from a single point and start out along equally separated directions may end up intersecting each other, due to the shape of the underlying network. A related complication is that rays emanating from a point uniformly may not cover the network uniformly. These and other issues are studied in [26].

An important issue that the works we reviewed do not address is the determination of the rate with which the truly optimal routes converge to the asymptotically optimal, optical routes as the number of nodes in the network increases. Fig. 4 gives us a sense of how fast this convergence is, but studying it formally is an important open issue.

IV. COOPERATIVE TRANSMISSIONS

The basic premise of cooperative transmissions [36] is that, instead of a single transmitter attempting to reach a receiver, multiple transmitter can combine their signals, and so increase their range. Likewise, multiple receivers can coordinate in receiving a transmitted signal (or signals), and so enhance their reception capabilities. In other words, nodes form virtual antenna arrays and so take advantage of all the benefits of multiple-input-multiple-output (MIMO) communication. The price we pay is complexity and the need for coordination among physically separate transceivers.

In the sequence of works [37], [38], [39], [40], [41] culminating in [42], the authors investigate the performance of cooperative transmissions in the massive wireless network limit, under a variety of settings. Here, we review some of their results. We note that there is no direct analogy between these results and any specific law of Physics. However, we include them here because of their affinity with the other results we review, and in particular their adoption of a macroscopic approach.

A. Basic Setting

Let us first examine the following simple setting considered in [42]: a set of nodes are distributed according to a uniform Poisson spatial process on a plane. There is a single source of data, placed at the origin, which must broadcast a packet to the rest of the nodes. All node receivers are susceptible to thermal noise, and receptions are successful if they exceed a given Signal to Noise Ratio (SNR) threshold \( \tau \). Transmitted power decays with distance \( l \) according to the equation \( P_{\text{received}}(l) = kl^{-2} \). Time is slotted and packet transmissions occupy a single slot.

At the beginning of the first slot, the source transmits the data packet. All nodes that are sufficiently close will decode the packet correctly, and will simultaneously transmit the packet again at the beginning of the second slot. The rest of the nodes will then receive a composite signal whose power equals, by hypothesis, the sum of the powers of all individual received signals. Those that receive the signal correctly will transmit it at the beginning of the third slot, and the process will be repeated, until either there are no new transmitters, or the data packet reaches the destination. Nodes that have transmitted the packet once do not transmit it again, even if they receive it correctly in a subsequent slot.

In Fig. 7 we have simulated the typical evolution of the packet propagation for different values of the SNR threshold. In the plots, nodes denoted by a filled circle transmit the packet during an odd slot. Nodes denoted by an empty circle transmit the packet during an even slot. Nodes denoted by an ‘x’ never decode the packet successfully, and so never transmit it. As the plots reveal, the evolution critically depends on the SNR threshold.

In particular, for low values of the SNR threshold, at each iteration the number of nodes that successfully decode the packet increases, and eventually the nodes successfully decoding the packet in a given iteration are placed along a ring whose width converges to a constant. An example is shown in Fig. 7(a). If we modify the situation by assuming that initially a large number of nodes that are placed in a circle centered at the origin hold the packet, then again at each iteration the nodes that successfully receive the packet are placed on a ring, whose width converges to the same constant. An example is shown in Fig 7(b). As the SNR threshold increases, simulations show that the successive rings become thinner and thinner, and eventually disappear, for a critical value of the SNR threshold. Above that critical value, irrespective of the number of nodes that initially possess the packet, after a few iterations the number of nodes that possess the packet will dwindle to zero. An example is shown in Fig 7(c).

The above observations reveal a behavior with a very interesting structure. However, the problem is very tough to deal with analytically, due to the randomness in the placement of nodes. The authors of [37] make the critical observation that the deadlock can be bypassed by assuming a massive network. In this context, the assumption means that the number of nodes in the plane goes to infinity, so that now there is a node in every location. As a result, we can describe the nodes that receive
the packet in the $i$-th iteration not in terms of their actual positions, which are microscopic quantities, but in terms of the two-dimensional set $S_i$ they occupy, which is a macroscopic quantity. This critically simplifies the problem, as it washes out randomness completely.

The authors proceed to show that the sets that transmit the packets in any iteration after the first few are approximately shaped like rings centered at the origin. In addition, the whole evolution of the system can be predicted in terms of a single function $h(x)$, which specifies the thickness of a ring given the thickness of the previous ring. In particular, if in the $k$-th iteration the width of the ring that transmits the packet is $r_k$, then in the next iteration the width will be $r_{k+1} = h(r_k)$. The authors prove that this function $h(\cdot)$ is increasing and concave. As a result, if $h'(0) < 1$ there is no positive solution to the equation $h(r^*) = r^*$ and necessarily the ring widths will converge to 0, and the fringes of the network will not be reached. If, however, $h'(0) > 1$, then there is a ring positive solution to the equation $h(r^*) = r^*$, and the width of the ring that transmits the packet in iteration $k$ will approach $r^*$, as $k \to \infty$. Both cases are plotted in Fig. 8. The phase transitions $h'(0) = 1$ occurs when the SNR threshold $\tau$ is

$$\tau = (\pi \ln 2) P_T \rho,$$

where $P_T$ is the transmit power of each relay, and $\rho$ is the relay density, measured in relays per $m^2$. For values of $\tau$ greater than the above, $h'(0) > 1$ and the transmissions do not die out. For values of $\tau$ smaller than the above, $h'(0) < 1$ and transmissions die out. These findings are consistent with the simulation results of Fig. 7.

**B. Extensions**

In the previous setting, if we allow nodes that successfully received the packet to retransmit it $m > 1$ times, as opposed to only one, then the situation remains approximately the same, with the only difference that the phase transition occurs for the new SNR threshold

$$\tau = [\pi \ln(m + 1)] P_T \rho.$$

Furthermore, if each node that has successfully received the packet in a previous iteration retransmits it in all future iterations, i.e., $m = \infty$, then there is no phase transition, and the packet always propagates through the whole network.

Until now, it was assumed that the channel was deterministic so that, given the location of all transmitters, the power received at the location of a receiver is not subject to any randomness caused by fading. In [37], [39], various fading models have been considered. If, for example, all transmitted signals occupy the same frequency band and arrive at the receiver with random phases, then the received signal exhibits Rayleigh fading. As a result, the nodes that transmit in the $i$-th iteration belong to sets $S_i$ that overlap. The propagation of the packet is now described by the new macroscopic parameter $P_k(x,y)$, which equals the probability that a node placed in $(x,y)$ will transmit the signal correctly in the $k$-th iteration. Again, if each node only transmits $m < \infty$
times, a phase transition exists: if the SNR threshold is set low enough, the sequence of probabilities \(P_k(x, y)\) resembles an expanding wave traveling arbitrarily far away from the source. If, however, the SNR threshold is high enough, then the probabilities \(P_k(x, y)\) resemble an expanding wave whose amplitude converges to zero as it moves away from the source.

Various other extensions have been considered, all taking full advantage of the massive network assumption. For example, in [40], [41] nodes are allowed to optimize their powers, depending on the stage when they transmit. In [38], the authors calculate the evolution of the packet error probability as packets move away from the source.

V. Load Balancing in Wireless Ad Hoc Networks

In the works we described until now, a single type of traffic was assumed, or at least a small number of them [21], each associated with its own set of sources and sinks. In the general case, however, each pair of nodes will be creating its own type of traffic. This case is considered, within the context of massive networks, in [10], [11], [43].

In more detail, the authors associate each pair of locations \(r_1\) and \(r_2\) in the network with a traffic demand density \(\lambda(r_1, r_2)\), measured in bps/m², and a unique path \(p\) that the packets will follow. The set of all paths over all destination pairs is denoted by \(P\). The traffic created over the whole network induces a flow of traffic which, at a location \(r\), and toward a direction \(\theta\), has an angular flux \(\Phi(P, r, \theta)\), measured in bps per meter per rad. The total volume of data that pass through \(r\) is described by the scalar flux \(\Phi(P, r) = \int_0^{2\pi} \Phi(P, r, \theta) d\theta\).

Note that, in line with the massive network assumption, the authors are interested in the data created by, or going through, or arriving at not a particular node, but a particular location. Similar formulations appear in the modeling of particle fluxes in Physics, notably in neutron transport theory. Within this framework, the authors are interested in determining the set of paths \(P_{\text{opt}}\) which minimizes the maximum scalar flux in the network:

\[P_{\text{opt}} = \arg\min_{P} \max_{r} \Phi(P, r).\]

In other words, the authors are interested in performing optimal load balancing, so that the most congested part of the network receives the minimum possible amount of traffic.

This optimization problem is sufficiently complicated so that the optimal routing strategy \(P_{\text{opt}}\) and the corresponding maximum value of the scalar flux \(\Phi_{\text{opt}}\) are hard to compute exactly, even in simple, highly symmetric networks. However, the authors manage to make a number of important contributions.

Firstly, they develop two lower bounds on the value of \(\Phi_{\text{opt}}\), blending massiveness arguments with standard tools from network theory, for example by partitioning the network in two parts and calculating the traffic that must flow through their boundary. Secondly, they develop a closed form expression for the angular flux at a location as a function of \(P\). Lastly, they focus on a highly symmetric topology, i.e., a network with the shape of a circular disk and with constant value of the traffic demand density and for that network develop routing strategies that result in an extremely flat traffic density across the whole disk.

In a more recent work [43], the authors use the same formulation to arrive at the unexpected result that, in the massive network limit, and for the purposes of performing traffic load balancing, single-hop routing is optimal. Therefore, no gain should be expected when multiple routes are employed. This is in stark contrast with fixed networks, where in general multi-path routing leads to reduced congestion.

It is interesting to note that a very similar setup was independently considered in [44], [45]. In this recently published work, the authors also attempt to minimize the maximum flow of traffic in a circular disk with uniform created traffic, within the class of rotationally symmetric flows. The authors show that this problem is isomorphic to the following problem: given a uniform distribution of source-destination pairs, and assuming that routes follow the laws of Optics, in particular (11), find what the refractive index should be, such that the maximum intensity of light is minimized. This problem can be solved by using an iterative algorithm and linear programming. The authors also propose a practical routing protocol termed Curveball Routing, that achieves results comparable to the optimum routing.

VI. Routing Using Potentials

In the previous sections we focused on massive wireless networks, and showed how the massiveness assumption can be used to derive useful, very tight analogies with various branches of Physics. However, a number of results have also appeared where analogies with Physics are made even in the context of finite sized networks. Such results are more practical, as they hold without taking the number of nodes to go to infinity. On the other hand, the analogies they use are not as tight, as networks with finite numbers of nodes are essentially discrete, and Physics typically deals with continuous media. In this section, we review a number of works that propose routing packets using potential functions similar to the potential function of Electrostatics.

Uncasting: In [46], the authors aim to build a routing protocol that optimally balances the following, potentially conflicting aims: on one hand, packets should follow relatively straight routes to their destination, in order to minimize the allocation of resources. On the other hand, packets should avoid going through congested regions, in order to minimize the delay to the destination, and also to minimize the wasting of resources, which typically is much higher in congested regions, due to collisions and retransmissions.

The proposed solution involves creating a potential field \(V(\cdot)\) that covers the network, so that each node \(n\) is associated with a given potential value \(V(n)\). The packet is then routed so that, in each hop, the node \(n_k\) that currently has the packet forwards it to that of his neighbors \(n_{k+1}\) with the minimum potential, assuming this minimum is smaller than the potential of \(n_k\). In other words, packets follow a discretized version of the path that a positive charge follows as it moves in an electrostatic field.

The potential function consists of two components, i.e., \(V(n) = V_1(n) + V_2(n)\). The first component \(V_1(n)\) resembles
a well, with a minimum located at the destination node \( n_d \). Various choices exist for the precise shape of the well, with the most obvious being \( V_1(n) = ad(n, n_d) + b \), where \( a \) and \( b \) are positive constants, \( n \) is an arbitrary node in the network, and \( d(n, n_d) \) is a distance metric between \( n \) and \( n_d \), such as the Euclidean Distance or the number of hops. The second component \( V_2(n) \) models the congestion in the network, and has a local hump in congested areas. For example, a reasonable choice if there is congestion around a single node \( n_c \) is \( V_2(n) = a \exp[-b \times d(n, n_c)] \), where \( a \) captures the intensity of the congestion, and \( b \) captures its geographical spread.

An example appears in Fig. 9. There, we consider a grid network of \( 40 \times 40 = 1600 \) nodes, with a single destination at the node with coordinates (20, 20), and four sources placed close to the corners of the network. Halfway between three of the sources and the destination are congested regions of varying levels of congestion intensity and spread. (For example, the congestion centered at the node \((10, 10)\) is much more localized than the congestion centered at the node \((10, 30)\).) As shown in the figure, a packet starting from the top right corner will follow a straight route to the destination, as there is no congestion to impede it. Packets starting from the other three corners, however, will deviate, in order to avoid the congested regions. The extend of the deviation depends on the extend of the congestion, as this is modeled by the hump in the potential function. In all four cases, the rule for choosing the next hop is the same: nodes pick their neighbor with the smallest potential.

This routing paradigm has a number of advantages: it is loop-free, it contains shortest-hop routing as a special case, it is greedy, and so simple to execute, it can adapt to the changing levels of congestion, and allows the designers many degrees of freedom in dealing with congestion. On the other hand, care must be exercised, to avoid the existence of local minima of the potential function, where a packet gets stuck, that do not coincide with the destination.

Anycasting: The work in [46] deals with unicasting, where each packet that is created has a single destination. However, the formulation can be easily extended to deal with anycasting, where each packet created must arrive at any one of a number of available destinations. This direction was extensively investigated in a sequence of papers culminating in the thesis [47].

In this work, as in [46], a potential field \( \phi(.) \) covers the network so that each node \( n \) is associated with a real number \( \phi(n) \), and the next hop is chosen according to the potentials of all neighbors. Here, however, the packet chooses the neighbor with the highest potential, provided it is higher than the potential of the current node. The potential is the superposition of a number of potentials \( \phi_j(n) \), each corresponding to a distinct destination \( j \) for the packet. In particular,

\[
\phi(n) = \sum_{j \in \mathcal{N}} \phi_j(n) = \sum_{j \in \mathcal{N}} \frac{Q_j}{d_j^k(n)}.
\]

In the above equation, \( \mathcal{N} \) is the set of anycast destinations. \( Q_j \) is the capacity of destination \( j \), and the larger it is, the more this particular destination will attract packets. \( d_j(n) \) is the distance, in hops, between destination \( j \) and the current node \( n \), and \( k \) is a positive real number. The larger \( k \) is, the fastest the potential decays as we move away from the destination.

An example appears in Fig. 10. There, we consider a grid network of \( 40 \times 40 = 1600 \) nodes, with three destinations placed at the grid locations \((10, 10)\), \((10, 30)\), and \((30, 20)\). The capacity of the third destination is double the capacity of the first two, and \( k = 0.5 \). In the figure, we also plot a number of routes, starting from various locations in the network. Note that the destination with the large capacity attracts packets from sources that are actually closer to the first two destinations, due to its increased capacity.

This anycasting scheme has a number of advantages. Firstly, the designers can optimize its performance in a particular application by tuning the capacities \( Q_i \), which specify how much...
each destination should attract packets, and the parameter $k$, which specifies how fast the attraction decreases with distance. Secondly, it can easily be deployed in a distributed fashion, by having each destination propagate its capacity and value of $k$ (which may be chosen to be different for each destination) through the whole network. Finally, as each packet can only leave a node to go to a node with a higher potential, it is loop free. On the other hand, similarly to the protocol in [46], care must be taken to avoid the existence of local potential maximums that are not destinations.

**Multipath Routing:** Using multiple routes between a source and a destination can be used for improving the reliability of the delivery, performing load balancing, and increasing the aggregate throughput of the communication. These benefits, however, materialize only if the routes are disjoint. In [48], the authors consider the problem of establishing multiple disjoint routes by using an analogy with electric fields. In particular, they assume that the source and the destination are aware of their own and the other party’s location. Assuming that a positive charge is placed in the source, and a negative charge is placed in the destination, an electric field is formed, with electric field lines starting from the source and ending at the destination. The authors propose the use of a few of these lines as routes. Obviously, packets can not follow the selected field lines exactly, as they have to be routed through the randomly deployed nodes in between. However, the packets should strive to select their next hops so that the resulting route will approximate the field line as accurately as possible. The lines selected should be sufficiently apart, to ensure that the resulting routes are disjoint. On the other hand, they should be chosen so that they are relatively short.

As elaborated in [48], this type of multipath routing has a number of very useful properties, inherited by its analogy to Electrostatics: The resulting paths are spatially disjoint without explicit coordination, the forwarding may be executed using only local information, the performance scales well in high mobility and heavy load environments, and the routing is simple and also very resilient against network failures and misbehaving nodes.

**Placement of Mobile Sinks in Wireless Sensor Networks:** In many applications of wireless sensor networks, it is highly desirable that the time until any of the nodes runs out of energy is maximized. Indeed, if a node runs out of battery, data forwarding will become harder in its vicinity and the environment around it may not be covered adequately. Therefore, it is best if the energy consumption is balanced among the nodes so that all of them run out of energy at roughly the same time.

Various different approaches have been considered that aim at balancing the energy consumption through the whole network. One method in particular involves the use of mobile sinks. The rationale is that if sinks can move, then the set of nodes that will be heavily involved in the forwarding of data, namely those nodes close to the sinks, will be rotating. Therefore, in the long run, the communication load will be balanced across all nodes and all nodes will remain operational for the maximum possible amount of time. The natural question to ask then, is how to rotate the locations where sinks should be placed.

In [49], the authors propose using an analogy with Electrostatics. In particular, sinks and nodes with low energy reserves are assigned positive charges, and nodes with high energy reserves are assigned negative charges. The sinks are left to move according to the combined electrostatic force they experience. As same-sign charges repel each other, and different-sign charges attract each other, the sinks will try to move far away from each other (and so cover the area uniformly), and also away from the sensors with low energy reserves. On the other hand, mobile sinks will be attracted by sensors with high energy reserves, who are willing to participate heavily in the collection of data. As energy levels of sensors will be changing, so will the charges associated with them, and so will the position of the mobile sinks. Studies by simulation show that this method outperforms all other approaches as far as balancing the energy consumption is concerned.

**VII. Routing Using Diffusion**

In Physics, diffusion is a generic term, which refers to the process under which matter or energy moves from locations of high concentration to locations of low concentration, so that the concentration becomes uniform. For example, a single drop of milk thrown in a glass of water is diffused in the whole glass, so that after a few moments the concentration of milk across the whole glass is constant. As another example, if we place a pot full of water on an electric stove, the heat produced in the heat element will be diffused across the whole pot.

A prominent characteristic of diffusion is that the rate of the diffusion is proportional to the gradient of the concentration. For example, heat flows faster between two locations as their temperature difference increases.

**Back-pressure in wireless multihop networks:** The first work where routing was performed by analogy to diffusion was the work in [50], although the connection with diffusion was not made at the time. In particular, the authors investigate the traffic carrying capabilities of queuing networks where time is slotted, various classes of traffic exist, each with its own sources and destinations, and with interdependent servers, where only certain combinations of the servers can be activated at any given time. The setup is sufficiently abstract to apply in many different contexts where there are interdependent servers, for example parallel computing systems, switches in routing, and wireless networks. In the case of wireless networks, servers correspond to wireless links, and since wireless transmissions interfere with each other, not all links can be active at the same time. The authors define a scheduling policy to be a rule for deciding, given the state of the buffers, which links will be activated in each slot, and which types of packets they will transmit. The authors also define the stability region to be all combinations of incoming traffic rates such that a scheduling policy exists so that buffer sizes do not explode.

Within this context, the authors determine the optimal scheduling policy, which stabilizes the traffic whenever this is possible, i.e., whenever the arrival rates fall within the stability region. The policy favors the transmission of packets across
links for which the transmitter has many more packets of the same type than the receiver. (Note that the destinations of a packet type never have a packet of that type in their buffers.) For example, if a node with many packets of some type has a neighbor who does not have any packets of that type in its buffer, the link connecting them will be frequently active, and there will be a large flow of traffic, in an attempt to balance out this buffer differential. Therefore, the larger the differential is in the concentration of packets across a link, the larger will be the flow of packets across the link. Therein lies an analogy with the diffusion of heat in materials: the greater the difference in temperature between two parts of a material, the faster will be the flow of heat energy from one part to the other, as nature strives to balance out differences and achieve an equilibrium. To conclude, the optimal scheduling policy replicates the behavior of nature when transferring heat.

**Diffusion in Wireless Sensor Networks**: A more recent work explicitly based on diffusion is [51], where the concept of Directed Diffusion in wireless sensor networks is introduced and evaluated. In this work, sinks interested in receiving sensor data of a given type diffuse interest packets across the whole network or parts of it (in which case the diffusion is directed). The diffused packets set up a gradient field, similar to the gradient field of the temperature in the diffusion of heat. However, in contrast to Physics, at each node there will be multiple gradients, one for every upstream node. Their numerical value can be associated with different things; for example, it could mean the rate with which the sink would like to hear updates of the requested data coming though that particular upstream neighbor, or the probability the particular upstream neighbor is chosen as the next hop of the data. Sensors with information matching the interest will use this gradient field to propagate information packets toward the sink. As the data sink starts collecting data from the sources, it may choose to modify the shape of the gradient field, in order to improve the efficiency of data gathering. The resulting flow of traffic resembles the diffusion of heat from regions with high temperature (where information is being created) to regions of low temperatures (where the information must arrive).

Another recent work, which also explicitly connects diffusion processes with the transportation of information, appears in [52]. There, the authors develop a detailed microscopic model of a sensor network, which incorporates the creation, transportation, and delivery of data, and then take the massive network limit to rigorously show that, macroscopically, the network can be described in terms of a quasi-linear diffusion-convection-reaction partial differential equation.

**Load Balancing in P2P Networks**: Dynamic load balancing methods inspired from diffusion processes have received significant attention in the context of distributed computing [53], [54], [55]. Here, we briefly discuss the more recent work in [56] on load balancing for peer-to-peer (P2P) file-sharing networks. Note that these networks are not wireless, however they have a number of important features that are shared by wireless networks, hence we mention the work in [56] in our study.

P2P networks consist of a number of nodes, all connected to a common underlying network, such as the Internet. The network can be described in terms of a graph, where the vertices of the graph are the nodes, and edges connect nodes that communicate directly with each other. Note that two nodes in the P2P network that are connected by an edge in the P2P graph might be several hops away in the underlying network. Some of their more popular applications are file sharing and P2P telephony.

In file-sharing P2P networks, each participating node is sharing its files with all other nodes, allowing the users to have transparent access to a potentially huge file system. In such systems, it is typical to use file replication, in order to minimize the access time. Therefore, any file can be replicated in a number of nodes other than the node where it was originally stored. A common problem in many replication algorithms is that some of the nodes in the network, typically those more centrally placed, are required to handle much more traffic and store many more file replicas than the nodes that lie in the periphery. Therefore, the need for load balancing arises.

The authors of [56] perform load balancing in a totally distributed manner by using an analogy with the diffusion of heat. In particular, each node $N$ is associated with a file replication ratio $L(N)$, which describes the volume of replications initiated until now by the node. The higher the ratio, the higher the load of the node. Ideally, we would like all nodes to have the same file replication ratio, and so share the burden of the network in a fair manner. To achieve this, each node periodically compares its ratio with the average of the ratios of its neighbors. If its ratio is higher than the average, the node becomes less willing to create a replicated file, and as a consequence its ratio will go down over time. The opposite will occur if the ratio is below the average: the node becomes more willing to replicate packets, and so its ratio will go up. The evolution of the system resembles the diffusion of heat in a metal plate, where the part of the file replication ratio is played by the temperature. There, a location with temperature higher than the average of the temperature of its surroundings will cool down and, conversely, a location with temperature lower than the average of its surroundings will heat up.

**VIII. Connectivity and Capacity of Ad Hoc Networks**

In chemistry and material science, percolation is the process by which a fluid passes through a porous material, such as a thin organic membrane or a slab of marble. Percolation theory is the branch of probability theory that studies various mathematical models of percolation and their properties, such as the conditions under which the fluid can flow through (i.e., percolate) the porous material model, and also the rate of the flow [57].

A typical model used in Percolation theory is the bond percolation model of Fig. 11: we are given a square grid of vertices, with each pair of neighboring vertices directly connected by an edge with probability $p$. A pair of vertices is indirectly connected if we can go from one vertex to the other by traveling along existing edges. Assuming that edges act as barriers to water, we would like to determine the possibility of
water seeping through the network. Assuming, for simplicity, that the network extends to infinity in both directions, it has been established that if $p > p_c = \frac{1}{2}$, then a cluster of connected vertices that extends to infinity in both directions exists almost surely, and as a result water can not seep through the network. If, however, $p < p_c$, then there is no cluster of infinite size, and water can seep through the network almost surely.

Another useful model is the continuum percolation model of Fig. 12. There, we place nodes randomly on the plane according to a Poisson process with density $\lambda$. Any two nodes are directly connected if they are within a distance $r$ of each other, and indirectly connected if we can go from one node to the other by jumping through directly connected nodes. In this context, we can imagine that the plane represents a swiss cheese, nodes represent holes in the cheese with radius $\frac{r}{2}$, and we would like to know if water can flow through. Percolation theory specifies that there is a critical density $\lambda_c > 0$, whose value we can only compute by simulation, such that if $\lambda > \lambda_c$, then an infinite sized cluster of connected nodes exists, and so water will go through the cheese. If, however, $\lambda < \lambda_c$, all clusters of connected nodes have finite sizes, and water can not flow through.

It has been well known for many years that certain Percolation theory models are intimately related to wireless network models (see [58] and references therein). For example, in the context of continuum percolation, if we assume that nodes are equipped with wireless transceivers with range $r$, then two nodes can communicate with each other through multiple hops if and only if they exist in the same cluster. Hence, the existence of a cluster of infinite size implies that nodes can communicate with nodes lying arbitrarily far away. Here, we review a number of recent results [59], [60], [61], [62] that are based on this connection.

In [59], authors use continuum percolation to study the connectivity properties of ad hoc networks, which consist exclusively of wireless nodes, and hybrid networks, which consist of wireless nodes and infrastructure nodes. In hybrid networks, infrastructure nodes are equipped with wireless transceivers but are also connected to each other through a secondary network, which can be used to support the communication of the wireless nodes. Two nodes are considered to be directly connected through the wireless medium if the distance between them is less than some global maximum $r$. Note that this condition for the connectivity totally disregards the fact that nodes may be subject to interference from neighboring nodes. The authors study both one-dimensional networks, such as those that are formed in highways and long and narrow alpine valleys, and two-dimensional networks, such as those formed in large uniformly populated cities.

In this setting, the authors show that percolation never occurs in one-dimensional networks, and so infrastructure nodes are necessary to ensure that wireless nodes that are far away from each other can communicate. In two-dimensional settings, if the node density exceeds a critical value $\lambda_c$, percolation occurs and so the wired infrastructure is not very useful, only helping a few isolated wireless nodes. If the node density is below the critical value, again the wired infrastructure is marginally helpful, unless a very high density of infrastructure nodes is used, as most wireless nodes will exist in isolated clusters with no connection to the infrastructure. The authors also consider the problem of bottlenecks, i.e., links that must carry a disproportionate large volume of traffic, due to the lack of other links connecting different parts of the network. To do this, they formally define a partition of the network in islands. For sub-critical densities, the islands are essentially disjoint clusters of finite size. For super-critical densities, the islands are relatively well connected areas of the network, connected to other islands through a few bridges, which represent the bottlenecks in the network.

In [59] two nodes can directly exchange packets as long as they are within some distance of each other. This is a greatly simplified model, that ignores the fact that transmissions typically compete with each other for the same bandwidth. In [60], a much more realistic model is assumed, under which two nodes can directly exchange packets as long as the Signal to
Interference and Noise Ratio (SINR) is above a given threshold $\beta$:

$$\frac{P_iL(x_i - x_j)}{N_0 + \gamma \sum_{k \neq i,j} P_kL(x_k - x_j)} > \beta. \quad (17)$$

In the above, $N_0$ is the thermal noise power, $P_i$ is the transmitter power of node $i$, $\gamma$ is a parameter which models the suppression of interference through the use of spread spectrum or other means, and $L(x_i - x_j)$ models the power attenuation between the locations $x_i$ and $x_j$ of nodes $i$ and $j$ respectively.

In this setting, an important question to ask is whether it is still possible for the network to percolate. The authors answer this question in the affirmative: as long as $\gamma$ and $\beta$ are sufficiently low, and $\lambda$ is sufficiently large, then the network percolates, and a node cluster of infinite size exists. The authors prove this by developing a discrete version of the network and employing results of bond percolation.

Percolation theory has also been used in research on the capacity of wireless networks. In particular, it has been used to sharpen, in two distinct and important ways, the results of Gupta and Kumar [9], which are the best known results to come out of this field in the last years. Gupta and Kumar have shown that, as the density $\lambda$ of the network approaches infinity, then the aggregate throughput available to the network increases with a rate between $\sqrt{\lambda}$ and $\sqrt{\frac{\lambda}{\log \lambda}}$. The result uses the model of (17) with the assumption that the attenuation function $L(l)$ is unbounded as the distance $l$ between the transmitter and the receiver goes to zero.

In [61], the authors remove the gap in the upper and lower bounds of the aggregate throughput, showing that, in fact, as the number of nodes goes to infinity, the aggregate throughput increases like $\sqrt{\lambda}$. They show this by constructing a discrete cellular version of the network, and selecting its operating parameters so that the network operates barely within the percolation regime. In particular, on the one hand communication is between nodes that are sufficiently close to ensure that spatial reuse is maximized, but on the other hand there is a sufficient number of connections between nodes to ensure that the resulting network can deliver an aggregate throughput on the order of $\sqrt{\lambda}$, without the formation of significant bottlenecks.

In [62], the authors show that if the attenuation function is substituted with a more realistic model that remains bounded as $l$ goes to zero, then the situation changes dramatically. In this case, in order to ensure that the network percolates, it is necessary for $\gamma$ to be decreasing at least as fast as $\frac{1}{\sqrt{\lambda}}$. In turn, this implies that the aggregate throughput will not be increasing at all with the node density, and the bounds predicted by Gupta and Kumar no longer hold.

IX. CONCLUSIONS

In this survey we review recent research activity on wireless networks that involves analogies with various branches of Physics. The results we present can be roughly partitioned into three types: in results of the first type, a massive network model is developed and shown to be analogous to a physical system, in the sense that both are governed by exactly the same set of equations. For example, it is shown that the optimal flow of traffic in a massive wireless sensor network resembles an electrostatic field, and that the optimal routes in a massive wireless network resemble rays of light. In results of the second type, we use tools developed in Physics to study networks that do not exactly map to any physical system. For example, we use Percolation theory to study the connectivity and capacity of networks, massiveness arguments to study the performance of cooperative transmissions schemes, and formulations developed in particle flux theory to determine the routing load. In results of the third type, protocols are developed that are inspired from, but only approximately follow, various physical systems. For example, a number of routing protocols are presented under which the flow of packets emulates the diffusion of heat or the move of test charges in electrostatic fields.

For many of these results the network is assumed massive, i.e., it consists of so many nodes that, in addition to the standard microscopic view, a macroscopic view starts to emerge. This view is based on the use of a small number of carefully selected macroscopic quantities. Jointly, these quantities retain just enough information to allow for a meaningful and tractable optimization of the network. In the context of wireless networks, this approach was proposed very recently, in [7] and independently in [8]. However, the wealth of results that have appeared since, by many different research groups, in many cases working independently of most of the others, demonstrate clearly its appeal and its power to provide deep insights in the operation of large wireless networks. It is hoped that this survey will raise the awareness of the wireless networking research community about these activities and so instigate more work along what promises to be a very fruitful research direction.

REFERENCES


