From the Physics of Interacting Polymers to Optimizing Routes on the London Underground

Chi Ho Yeung [*], **David Saad** [*], **and K. Y. Michael Wong** [†]

[*]The Nonlinearity and Complexity Research Group, Aston University, Birmingham B4 7ET, United Kingdom, and [†]Department of Physics, The Hong Kong University of Science and Technology, Hong Kong

Submitted to Proceedings of the National Academy of Sciences of the United States of America

Optimizing paths on networks is crucial for many applications, from subway traffic to Internet communication. As global path optimization that takes account of all path-choices simultaneously is computationally hard, most existing routing algorithms optimize paths individually, thus providing sub-optimal solutions. We employ the physics of interacting polymers and disordered systems to analyze macroscopic properties of generic path-optimization problems and derive a simple, principled, generic and distributed routing algorithm capable of considering simultaneously all individual path choices. We demonstrate the efficacy of the new algorithm by applying it to: (i) random graphs resembling Internet overlay networks; (ii) travel on the London underground network based on Oyster-card data; and (iii) the global airport network. Analytically derived macroscopic properties give rise to insightful new routing phenomena, including phase transitions and scaling laws, which facilitate better understanding of the appropriate operational regimes and their limitations that are difficult to obtain otherwise.

Routing | Optimization | Transporation Networks | Communition Networks | Disordered Systems | Polymers

Introduction

Path optimization affects many of our daily activities. While much attention has been dedicated to routing algorithms for Internet applications such as instant messengers and peer-to-peer systems [1, 2], many other essential routing applications have attracted less attention; from water distribution networks [3], sensor networks [4], military convoy movements [5] to journey planners [6, 7]. In many applications, enormous costs are incurred due to traffic congestion or non-essential and redundant capacity. Due to the computational costs involved, most existing routing algorithms are static and based on selfish decisions, with non-adaptive routing tables indicating the shortest path to destinations regardless of local traffic [8, 9]. Dynamic routing protocols do exist, but they are either heuristic, probabilistic or insensitive to other individual path decisions which dynamically constitutes the traffic [10, 11]. A more global approach that takes into account all individual path decisions is crucial for efficient use of overstretched infrastructure. For instance, one may suppress congestion by minimizing overlaps with other routes, or decrease the number of active nodes by consolidating paths to reduce infrastructure demands or energy consumption. The latter is particularly important in the context of the Internet as it can consume up to 4% of the electricity generated [12]. Future applications include individualized routing and optimal resource management of pre-booked air and road traffic.

The difficulty in deriving a globally-optimal algorithm, in contrast to greedy local ones, lies in the simultaneous assignment of multiple interacting paths to minimize a global cost, as the optimal path between any particular source-destination pair depends on all other paths choices. Such interaction is highly non-local, as paths between different source-destination pairs may partially overlap. Existing algorithms either ignore these interactions [8, 9], or use heuristics to approximate them [10, 11]; both approaches result in sub-optimal solutions. A substantial effort has been devoted to the development of highly efficient routing methods, for instance multi-commodity flow algorithms [13, 14, 15, 16, 17, 18, 19]. However, most methods are based on weighted linear objective functions and real variables and aim specifically at satisfying capacity constraints; they have limited flexibility in addressing the variety of non-linear cost functions one may want to optimize in different scenarios, especially concave costs and integer variables. A more detailed discussion is provided in Section S4 of the *SI Appendix*.

Here we employ statistical physics-based methods used in the study of interacting polymers [20] and spin glasses [21, 22] to obtain both a macroscopic description of the routing problem and microscopic solutions for given instances; the latter leads to a simple, generic and distributed routing optimization algorithm. The algorithm resembles message passing techniques that have been developed independently in a number of disciplines [21, 23, 24] and have been successfully applied to a variety of problems from prototyping [25] to solving hard computational problems [26] and control of complex systems [27]. Here we demonstrate the potential and efficacy of our routing algorithm by applying it to random networks, individualized routing on the London subway network and the global airport network. Together with other benchmark tests described in the *SI Appendix*, we demonstrate that our algorithm achieves better optimization compared to existing heuristics and state-of-the-art approximation algorithms in various routing scenarios; moreover, it is distributed, principled and does not require fine-tuning of free parameters.

In addition to the significant algorithmic advances, several macroscopic phenomena including a phase transition, scaling rules as a function of network size and non-monotonic growth in mean path length as a function of traffic volume are revealed; these cannot be obtained by numerical studies and provide new insights and understanding of optimal routing on sparse networks.

Model

Consider a system of M polymers interacting on a network of N nodes. Each node $i = 1, \ldots, N$ is connected to k_i neighbors denoted by the set $\mathcal{L}_i$ and the connectivity matrix $A_{ij} = A_{ji} = 1$ when i and j are connected and zero

Reserved for Publication Footnotes

otherwise. Each polymer $\nu=1,\ldots,M$ has two fixed ends and occupies a path described by a self-avoiding walk on the network, i.e. consecutive segments occupy topological neighbors and each polymer ν goes through a node at most once. We denote the variable $\sigma_i^\nu = 1$ when polymer ν occupies node i and $\sigma_i^\nu = 0$ otherwise, and the number of polymers occupying i as $I_i = \sum_\nu \sigma_i^\nu$. To penalize or encourage polymer overlap, we define the Hamiltonian $\mathcal{H}$ to be a non-linear function of the normalized flow $\lambda_i = I_i/M$, namely

$$\mathcal{H} = M \sum_i \phi(\lambda_i). \qquad [1]$$

The analytic solution and derived algorithm are generic for any ϕ. While the current framework focuses on undirected polymers and costs which incur at vertices, it is clear that in some applications costs incur at the edges and edges may be directed and weighted. Our framework, derivation and algorithms accommodate costs on edges (using a factor graph representation) as well as directed and weighted polymers, making them suitable for most routing scenarios. The derivation and corresponding algorithms are given in *SI Appendix* Sec. S3. We would like to point out that the algorithm presented below already accommodates directed traffic.

This model is equivalent to a setting of M source-destination pairs, which we term *communications*, each of which occupies a path on a network with N nodes. The variable λ_i is thus the normalized traffic on node i and $\mathcal{H}$ the corresponding cost function. In the physical framework and the zero-temperature limit, we minimize $\mathcal{H}$ to obtain the ground state of the system or the optimal path configuration of the corresponding routing system. Some simple forms of $\mathcal{H}$ are already meaningful, for instance $\phi(x) = x^\gamma$, where the cases with $\gamma > 1$ penalize overlaps to suppress congestion while $\gamma < 1$ encourages overlaps to aggregate traffic [28, 29, 30]. The case of $\gamma = 1$ reduces to $\mathcal{H} \propto \sum_\nu (\sum_i \sigma_i^\nu)$ whose ground state corresponds to shortest-path routing.

Methods

Theoretical Approach. The main obstacle in accounting for the interaction between paths is in keeping track of the cost at local nodes or edges while maintaining path-integrity between the two end points and avoiding redundant loops. Therefore, in addition to the cost at the various nodes, given by Eq. (1), we introduced a technique used in polymer physics [20], in the study of self-avoiding walks [31, 32], to enforce the appropriate path constraints.

The method is based on representing each node as an n-component vector $\vec{S}$ of length $\sqrt{n}$. Denoting the angular integration over $\vec{S}$ as $\int_\odot d\vec{S}$, it has been shown [20] that all positive moments of S_a vanish in the limit $n \to 0$ except the second moment $\frac{1}{C_n}\int_\odot d\vec{S} S_a^2 = 1$ for any component a in $\vec{S}$, where $C_n = \int_\odot d\vec{S}$ is a normalization constant. It then implies that when $n \to 0$ all nonvanishing terms that contribute in

$$\prod_{i=1}^{N} \left(\frac{1}{C_n} \int_\odot d\vec{S}_i \right) S_{x,a} S_{y,a} \prod_{(kl)} \left(1 + A_{kl} \vec{S}_k \cdot \vec{S}_l \right) \qquad [2]$$

are of the form $A_{xk_1} A_{k_1 k_2} \cdots A_{k_l y} S_{x,a}^2 S_{k_1,a}^2 S_{k_2,a}^2 \cdots S_{k_l,a}^2 S_{y,a}^2$, where k_i represents the i-th node index of the corresponding path/polymer segment; these sequences represent self-avoiding paths over nodes $(x, k_1, k_2, \ldots, k_l, y)$, joining the end nodes x and y [20]. Each node that is part of these paths incurs a cost as in Eq. (1); a sum over all possible paths of all communications provides the partition function $\mathcal{Z}$, as detailed in *SI Appendix* Sec. S1.1. To obtain typical macroscopic properties one needs to average $\mathcal{Z}$ over topologies (given a degree distribution) and node-pair choices, termed *quenched disorders* in statistical physics. This requires the use of the replica or cavity methods of spin glass theory [21, 22], as presented in *SI Appendix* Sec. S1.

The aim of the analysis is two-fold: (1) At the macroscopic level, we derive the stable traffic distribution $P(I)$ in the limit of very large systems to obtain the average cost (energy) $\langle E \rangle = \langle \phi(I/M) \rangle$, the average path length, given by the total occupancy divided by M, i.e. $\langle L \rangle = \frac{N}{M}\langle I \rangle$, and the average fraction of idle nodes given by $f_{\text{idle}} = \langle \delta(I) \rangle$, detailed in *SI Appendix* Sec. S1.4. Angled brackets denote an average over $P(I)$, which includes averages over all variable states for a given network and over choices of network and end-point instances. (2) At the microscopic level, the cavity based analysis [33] translates to an algorithm which optimizes path configuration in a principled, distributed and computationally efficient manner.

Optimization algorithm. The analytical solutions for infinite systems translate into an optimization algorithm valid for finite systems, as detailed in *SI Appendix* Sec. S2. The derived algorithm is based on sending a couple of messages $a^\nu_{j\to i}$ and $b^\nu_{j\to i}$ at the zero temperature limit, from node j to node i for each index ν; these characterize the energy contributions of communication ν at edge $j \to i$, originated from the source and destination directions, respectively. The messages take the form:

$$a^\nu_{j\to i} = \begin{cases} \min\limits_{l\in\mathcal{L}_j\setminus\{i\}} \left[a^\nu_{l\to j}\right] - \min\left[-\phi'(\lambda_j^{\nu*}), \min\limits_{\substack{l,r\in\mathcal{L}_j\setminus\{i\}\\ l\neq r}} \left[a^\nu_{l\to j} + b^\nu_{r\to j}\right]\right], & \Lambda_j^\nu = 0 \\ -\min\limits_{l\in\mathcal{L}_j\setminus\{i\}} \left[b^\nu_{l\to j}\right], & \Lambda_j^\nu = 1 \\ \infty, & \Lambda_j^\nu = -1 \end{cases} \qquad [3]$$

$$b^\nu_{j\to i} = \begin{cases} \min\limits_{l\in\mathcal{L}_j\setminus\{i\}} \left[b^\nu_{l\to j}\right] - \min\left[-\phi'(\lambda_j^{\nu*}), \min\limits_{\substack{l,r\in\mathcal{L}_j\setminus\{i\}\\ l\neq r}} \left[a^\nu_{l\to j} + b^\nu_{r\to j}\right]\right], & \Lambda_j^\nu = 0 \\ \infty, & \Lambda_j^\nu = 1 \\ 0, & \Lambda_j^\nu = -1 \end{cases} \qquad [4]$$

where $\Lambda_j^\nu = +1, -1$ for source and destination, respectively, and is zero otherwise; the general cost function ϕ and the set of nodes in the neighborhood of node j is denoted as $\mathcal{L}_j$. The value of $\lambda_j^{\nu*}$ is given by the solution of λ_j^ν in

$$\lambda_j^\nu = \frac{1}{M} + \frac{1}{M} \sum_{\mu\neq\nu} \Bigg\{ |\Lambda_\mu| + (1 - |\Lambda_\mu|)\Theta\left(-\phi'(\lambda_j^\nu) - \min_{\substack{l,r\in\mathcal{L}_j\\ l\neq r}} \left[a^\mu_{l\to j} + b^\mu_{r\to j}\right] \right) \Bigg\}, \qquad [5]$$

The step function $\Theta(x)$ takes values $\Theta(x) = 0, 0.5, 1$ for $x < 0, x = 0$ and $x > 0$, respectively. Solutions of Eq. (5) are obtained by setting $\lambda_i^\nu = I/M$ and a test integer I starting from $I = 0$ until a self-consistent λ_j^ν is found. Finally,

after the set of messages in Eqs. (3) and (4) converges to non-fluctuating values, the optimal configuration of path ν on each node j is given by

$$\sigma_j^\nu = |\Lambda_\nu| + (1 - |\Lambda_\nu|)\Theta\left(-\phi'\left(\lambda_j^{\nu *}\right) - \min_{\substack{l,r\in\mathcal{L}_j \\ l\neq r}}\left[a_{l\to j}^\nu + b_{r\to j}^\nu\right]\right), \quad [6]$$

where $\lambda_j^{\nu *}$ is the solution of Eq. (5) after convergence, and $\sigma_j^\nu = 1$ if the communication ν passes through node j and zero otherwise. The generalized algorithms which accommodate weighted and directed communications, generic costs on nodes and edges, as well as separate costs defined on directed edges are given in *SI Appendix* Sec. S3. The computational complexities of these algorithms are discussed in *SI Appendix* Sec. S2.2.

In some instances the iterative equations fail to converge, this suggests that solution space in the infinite system case is fragmented and non-ergodic; this corresponds to*replica symmetry breaking* (RSB) [21, 22],a complicated energy landscape with numerous local minima that typically hinder algorithmic convergence (details in *SI Appendix* Sec. S5). This is typical in the case of hard computational problems. Convergence is improved by assigning a random bias ϵ_i to each node [34], akin to an external field, guiding the system to one of the local minima. These biases can be easily incorporated in the present formulism by replacing $\phi(x)$ with $\phi_i(x)$ for each node i such that $\phi_i(x) = \phi(x) + x\epsilon_i$. In cases where a large number of source-destination pairs are identical, we further replace ϵ_i by ϵ_i^ν for each communication ν to break the degeneracy brought about by Eq. (5). Details can be found in *SI Appendix* Sec. S2.1.

Results

Microscopic Solution - Finding Best Paths. Employing the suggested algorithm, we can optimize path choices using the cost $\mathcal{H} \propto \sum_i I_i{}^\gamma$. We illustrate the characteristic results obtained by applying the algorithm using two costs, with $\gamma = 2$ (convex, $\gamma > 1$) and $\gamma = 0.5$ (concave, $\gamma < 1$), to a system of 10 source-destination pairs communicating on a random regular graph with $N = 50$ and $k = 3$ as shown in Fig. 1.

Figure 1(a) demonstrates how a cost with $\gamma > 1$ penalizes congestion: the blue, orange and violet communications are routed via non-shortest paths to avoid overlap, especially in the central congested part of the network. This holds when traffic is heavy and one aims to distribute it uniformly. In contrast to the reduced-congestion solutions, Fig. 1(b) shows solutions obtained for $\mathcal{H} \propto \sum_i I_i{}^{0.5}$, aimed at concentrating traffic. More specifically, the blue, orange and violet communications are all routed via the central congested part of the network that mainly consists of source and destination nodes, making best use of these nodes as relays and leaves many of the other nodes idle. In the case of the Internet or transportation networks, idle nodes can be switched off to save resources.

To demonstrate the efficacy of the algorithm for more realistic systems we examine the performance of the algorithm on the London subway network based on real passenger source-destination data obtained by the Oyster card system [35]. We report results for vertex costs only, but similar pictures have been obtained for edge costs and directed traffic. Figure 3(a) shows how congestion is reduced by the algorithm when $\gamma = 2$ is used and traffic is fairly uniform even in the central region (see inset), at the cost of longer individual routes for global optimization. Table 1 shows that the cost $E = \sum_i I_i^2$ obtained by our algorithm is 20.5% smaller than that of the shortest path configuration obtained by the commonly used Dijkstra algorithm [9], with only a slight increase in average path length by 5.8%. Practically, traffic optimization of this type may be achieved through differential pricing, or by auxiliary information provided either individually or globally. On the other hand, when $\gamma = 0.5$ is used, paths for the same passenger set are consolidated at major routes and stations as shown in Fig. 3(b). While the size of some of the nodes increases, other branches such as the ones passing through "Holborn" and "Great Portland Street" (see inset) are all but idle. This scenario may be relevant at times when the service is reduced for some reason, for instance a strike or at late evening; service on the shared branches can remain active while the frequency of other less-loaded services decreases.

To better compare the solutions obtained in the two scenarios, we plotted the corresponding traffic at individual stations for the London underground data set in descending order (for $\gamma = 0.5$) as shown in the inset of Fig. 2. The optimized states of $\gamma = 2$ show less traffic for overloaded stations and higher traffic for less-loaded ones; for instance, "Green Park".

Similar experiments were carried out on the global airport network [36]. Applying the optimization algorithm (3-4) to the data one obtains the results presented in Fig. 2 and Fig. 5. Similar trends to those of the subway network are observed: air-traffic consolidates at airports that are on main routes in the case of $\gamma = 0.5$, such as Frankfurt, Toronto and Beijing; while several popular airports such as Tokyo, Newark and Hong Kong show a reduced air-traffic in the case of $\gamma = 0.5$ represented by the red line. Table 1 shows the cost obtained by our algorithm when $\gamma = 2$ is 56% lower than

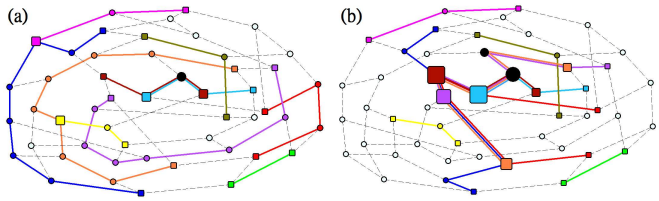

Fig. 1. Optimized path configurations on a regular random network. The network comprises 50 nodes (each with $k = 3$) and 10 source-destination pairs. The corresponding costs are (a) $\mathcal{H} \propto \sum_i I_i{}^2$, and (b) $\mathcal{H} \propto \sum_i I_i{}^{0.5}$. The path of each communication is illustrated by nodes and edges of a specific color, while black nodes are shared by more than one path. The size of a node is proportional to the amount of traffic through it, and square nodes represent source or destination of each communication.

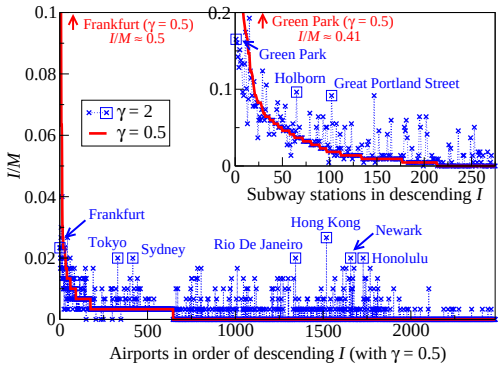

Fig. 2. Optimized traffic at individual airports and London subway stations (inset). The airports and stations are plotted in descending order of traffic in the optimized state of $\mathcal{H} \propto \sum_i I_i{}^{0.5}$ (red lines). Symbols (×) in blue correspond to the optimized traffic with $\mathcal{H} \propto \sum_i I_i{}^2$. Squared symbols refer to airports and stations mentioned in the text that are much higher than the red lines. The optimized airport traffic is obtained from the single instance shown in Fig. 5 and the optimized subway traffic is obtained by averaging over the 30 passenger sets as in Table 1.

	$\gamma = 2$		$\gamma = 0.5$	
	$\frac{E_{\rm P}-E_{\rm D}}{E_{\rm D}}$	$\frac{L_{\rm P}-L_{\rm D}}{L_{\rm D}}$	$\frac{E_{\rm P}-E_{\rm D}}{E_{\rm D}}$	$\frac{L_{\rm P}-L_{\rm D}}{L_{\rm D}}$
London subway network	$-20.5 \pm 0.5\%$	$+5.8 \pm 0.1\%$	$-4.0 \pm 0.1\%$	$+5.8 \pm 0.3\%$
Global airport network	$-56.0 \pm 2.0\%$	$+6.2 \pm 0.2\%$	$-9.5 \pm 0.2\%$	$+8.6 \pm 1.2\%$
	$\frac{E_{\rm P}-E_{\rm MC}(\alpha^*)}{E_{\rm MC}(\alpha^*)}$	$\frac{L_{\rm P}-L_{\rm MC}(\alpha^*)}{L_{\rm MC}(\alpha^*)}$	$\frac{E_{\rm P}-E_{\rm MC}(\alpha^*)}{E_{\rm MC}(\alpha^*)}$	$\frac{L_{\rm P}-L_{\rm MC}(\alpha^*)}{L_{\rm MC}(\alpha^*)}$
London subway network	$-0.70 \pm 0.04\%$	$+0.72 \pm 0.10\%$	No existing algorithm	
Global airport network	$-3.09 \pm 0.59\%$	$+0.90 \pm 0.64\%$	for comparsion	

Table 1. A comparison of average cost $E = \sum_i I_i^\gamma$ and path length $L = \frac{1}{M}\sum_i I_i$ obtained by our algorithm (P), the Dijkstra algorithm (D) and the modified min-cap congestion aware algorithm (MC) [13] at individual optimal α^* for each instance. Results are averaged over sets of source-destination pairs recorded in each 1 minute interval between 8:30 am – 9:00 am on one Wednesday in November 2009 for the London subway network, and 5 sets of 300 randomly drawn source-destination pairs for the global airport network. The values after the $\pm$ signs indicate to the corresponding standard error.

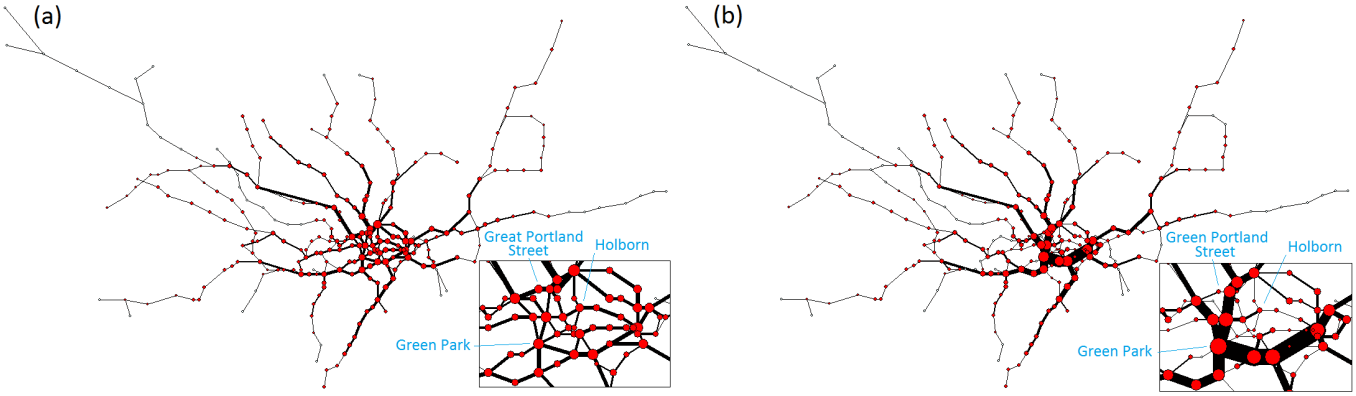

Fig. 3. Optimized traffic on the London subway network. A total of 218 real passenger source-destination pairs are optimized, corresponding to 5% of data recorded by the Oyster card system between 8:30am - 8:31am on one Wednesday in November 2009 [35]. The network consists of 275 stations. The corresponding costs are (a) $\mathcal{H} \propto \sum_i I_i^2$, and (b) $\mathcal{H} \propto \sum_i I_i^{0.5}$. Red nodes correspond to stations with non-zero traffic. The size of each node and the thickness of each edge are proportional to traffic through them. Insets: zoom on the central region. Nodes are drawn according to their geographic position.

that obtained by the Dijkstra's shortest paths, with a slight increase in path length of 6.2%. This may be due to the availability of a large number of alternative paths in airport network. We note that a lower cost is also achieved in the cases of $\gamma = 0.5$. These results show that our algorithm optimizes a given generic cost, at a price of modest increase in the average path length.

To evaluate the performance of the suggested algorithm (with $\gamma = 2$ only) we compared our results against those obtained using a representative state-of-the-art congestion-aware routing algorithm, which we call the *min-cap* (MC) algorithm [13], based on multi-commodity flow. As the latter aims to optimize a linear cost, we have introduced a tunable parameter α such that the quadratic cost is optimized by an extensive search for an optimal α^* (see *SI Appendix* Fig. S8). Details are found in *SI Appendix* Sec. S4. We emphasize that *this comparison is limited to congestion-aware algorithms* ($\gamma \geq 1$) as we have not identified existing efficient optimization algorithms for concave costs that facilitate route consolidation, e.g. the results shown in Figs. 1(b), 3(b) and 5(b).

Table 1 shows a modest gain in cost over the optimized MC results at individual α^* for each run, far less than the gain obtained with respect to Dijkstra's algorithm. Nevertheless, our algorithm provides a lower energy for all α values, unachievable by the MC algorithm even after fine-tuning (see *SI Appendix* Fig. S8). Our algorithm also results in shorter average path length L in addition to lower cost E in random regular graphs (see *SI Appendix* Table S1), used as a controlled benchmark problem. Moreover, it is distributed, principled, does not require fine-tuning of free parameters and, most importantly, has the flexibility to accommodate any (non-pathological) cost function designed to address specific needs.

Path Adaptivity. Figure 4 illustrates the adaptivity of our algorithm after removing the London subway station "Bank" (black node). Nodes and edges which show an increase (decrease) in optimized traffic are colored red (blue), respectively, with their size and thickness proportional to the magnitude of

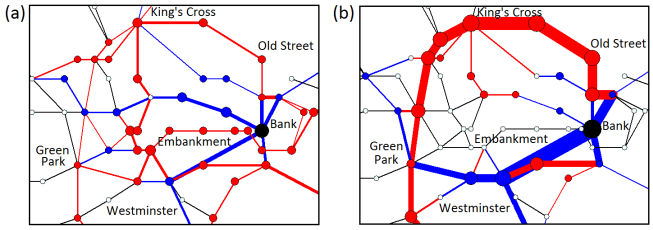

Fig. 4. Changes in optimized traffic in the central London subway network after the removal of the station "Bank" (black node). The corresponding costs are (a) $\gamma = 2$ and (b) $\gamma = 0.5$. Nodes and edges which show an increase (decrease) in traffic appear in red (blue), where their size and thickness correspond to the magnitude of increase (decrease). Nodes and edges with no traffic changes appear in white and black, respectively. Passengers source-destination pairs are identical to those of Fig. 3, except for the removal of pairs starting or ending destinations in "Bank".

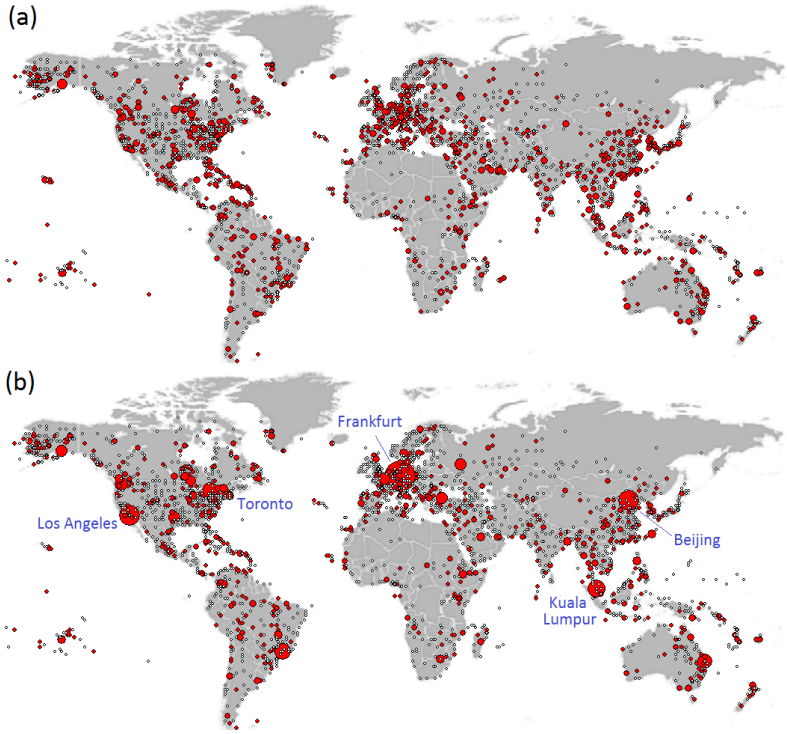

Fig. 5. Optimized traffic at individual airports of the global air network. A total of 2480 airports constitute nodes while edges represent the existence of direct flights between airport pairs [36]. Since the real demand in terms of source-destination pairs is unavailable, it was artificially generated by selecting a set of 300 randomly drawn source-destination pairs. Red nodes correspond to airports with non-zero traffic; the size of nodes indicates the air-traffic through particular airports, edges are omitted for clarity. (a) For $\mathcal{H} \propto \sum_i {I_i}^2$ traffic is routed to be almost uniformly distributed to reduce congestion. (b) For $\mathcal{H} \propto \sum_i {I_i}^{0.5}$ air-traffic consolidates at the main hubs.

increase (decrease). Nodes and edges with no traffic changes are in white and black, respectively. In the case of optimization using $\gamma = 2$, the original traffic through "Bank" is re-routed either via "Embankment" or via "Old Street". This re-distribution of traffic cannot be achieved by ordinary algorithms such as routing tables, shortest-path or minimal weight routing without taking into account the interaction between paths.

On the other hand, in the case of $\gamma = 0.5$, almost all the original traffic through "Bank" is diverted to "Old Street". As the original traffic via "Bank" is substantial (see inset of Fig. 3(b)), significant changes at some stations have to be made, although only a small number of stations are subject to re-routing compared to the case of $\gamma = 2$.

Macroscopic Behavior in Routing. In addition to the microscopic solutions obtained, we would like to explore the macroscopic behavior of the system. We first examine the dependence of average path length $\langle L \rangle$ on the number of interacting communications M. Random regular networks, Erdös-Rényi (ER) and scale-free (SF) graphs are studied as they serve as standard benchmark problems and resemble overlay networks on the Internet. Theoretical results are obtained by solving numerically a set of recursive equations described in *SI Appendix* Sec. S1.4; simulation results are obtained using Eqs. (3) and (4). The inset of Fig. 6(a) shows results obtained for random regular graphs. Two remarkable phenomena are observed for both $\gamma = 2$ or $\gamma = 0.5$: (i) average path length $\langle L \rangle$ peaks at intermediate M instead of increasing monotonously; (ii) it approaches asymptotically the shortest path L_1 as $M \to \infty$ (formally, the value of $\langle L \rangle$ when $M = 1$). Small deviations between theory and simulations are due to finite size effects.

The observed non-monotonic trends imply the existence of interesting routing phenomena. In the case of $\gamma = 2$, it implies that the system is very sensitive to congestion in the intermediate range of M. Particularly when M is small, many communications are routed through longer routes as they face stiff competition for shorter ones. However, as M increases further, traffic become more homogeneous and $\langle L \rangle$ decreases since communications are routed via shorter routes as longer ones are equally congested, matching the experience of frustrated drivers on congested roads. This is reflected in the lower cost obtained by our algorithm in comparison with Dijkstra algorithm, which peaks at 20% for intermediate M as shown in Fig. 6(b). Similar trend is observed for $\gamma = 0.5$ as different communications co-operate to share routes in the intermediate range of M. As M increases further, traffic becomes more homogeneous and there is less advantage to prefer

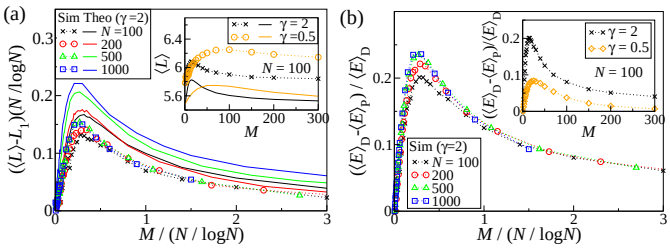

Fig. 6. Dependence of the optimized state on the number of communications. (a) The rescaled path length $(\langle L \rangle - L_1)(N/\log N)$ and (b) the cost difference $(\langle E \rangle_D - \langle E \rangle_P)/\langle E \rangle_D$ (D and P stand for the Dijkstra algorithm and our algorithm respectively), as a function of the rescaled number of communications $M/(N/\log N)$, for random regular graphs with $N = 100, 200, 500, 1000$ and $k = 3$; results were obtained for $\mathcal{H} \propto \sum_i {I_i}^2$. The value of L_1 in (a) corresponds to the value of the shortest path $\langle L \rangle$. Insets: (a) $\langle L \rangle$ and (b) $(\langle E \rangle_D - \langle E \rangle_P)/\langle E \rangle_D$ as a function of M for $N = 100$ on random regular graphs of degree $k = 3$, with cost exponents $\gamma = 2$ and $\gamma = 0.5$. The error bars for simulation results are of the order of the symbol size. All simulation results are averaged over 2000 realizations.

a busy but longer route, making shorter routes more cost-effective. We note that the peak in $\langle L \rangle$ for the case of $\gamma = 2$ occurs at a smaller M value compared to $\gamma = 0.5$, implying that traffic homogeneity is achieved at smaller M in the case of $\gamma = 2$.

While similar behaviors are observed for ER graphs (see *SI Appendix* Sec. S6), SF networks show a much slower decrease of $\langle L \rangle$ after attaining its maximum, possibly due to the intrinsic node degree inhomogeneity which leads to traffic inhomogeneity even at large M. This suggests that shortest-path routing is effective when M is large and topology is homogeneous, but not in networks with high degree variability.

The scaling property of path lengths is shown in Fig. 6(a). Rescaled path length $(\langle L \rangle - L_1)(N/\log N)$ with $\gamma = 2$ at system sizes $N = 100, 200, 500$ and 1000, plotted as a function of the rescaled number of communication, $M/(N/\log N)$ fall on top of each other almost identically. A similar data collapse is also observed in ER graphs shown in *SI Appendix* Sec. S4. It implies that the non-monotonic behavior observed for path lengths, and thus the network sensitivity to congestion, depend on M and N only through $M/(N/\log N)$. The latter is roughly proportional to the average traffic on a node since $\log N$ is proportional to the average shortest distance between any two nodes in random regular networks [37, 38] and ER graphs [39]. In other words, the optimal behavior of routing on these graphs depends only on the average node traffic, regardless of system size and number of communications. The rescaling also appears in the reduced cost obtained by our algorithm as shown in Fig. 6(b). Note that theoretical results have been obtained in the infinite system limit; finite N values have been introduced here merely to determine the scaling properties of M.

We have also examined the fraction of idle node as a function of γ. This revealed a phase transition, an abrupt change in the fraction of idle nodes around the $\gamma = 1$ value (see *SI Appendix* Fig. S11 and Sec. S7). The implication is that even a small change in the power γ is sufficient to effectively power down unnecessary routers or close redundant subway stations, with little impact on the cost or average route length.

Discussion

Optimal routing is one of many hard problems on networks that one should tackle in order to use limited and usually overstretched resources efficiently. The common characteristic of these problems is their global nature and thus the difficulty in solving them at both macroscopic and microscopic levels with limited computational resources. By applying methods from the physics of interacting polymers and disordered systems we obtained typical properties of routing problems and derive a readily applicable, principled, generic, distributed and adaptive routing algorithm. Improvements over state-of-the-art algorithms in the intermediate traffic regime where $M \sim N \log N$ are considerable but are modest in the very sparse and dense traffic regimes. These findings will have direct impact on a number of different research areas of practical and societal relevance, from traffic to communication and logistics; but more importantly, may open the way for solving many other crucial and non-localized problems on networks.

ACKNOWLEDGMENTS. This work is supported by EU FET project STAMINA (FP7-265496), Royal Society Exchange Grant IE110151 and Research Grants Council of Hong Kong (605010).

1. Huitema C (1995) Routing in the Internet (Prentice Hall, Englewood Cliffs, NJ).
2. Moy JT (1998) OSPF: Anatomy of an Internet Routing Protocol (Addison-Wesley, Reading, MA).
3. Vasan A, Simonovic SP (2010) Optimization of water distribution network design using differential evolution. Journal of Water Resources Planning and Management 136:279–287.
4. Rangwala S, Gummadi R, Govindan R, Psounis K (2006) Interference-aware fair rate control in wireless sensor networks. Computer Communication Review 36:63–74.
5. Chardaire P, McKeown GP, Verity-Harrison SA, Richardson SB (2005) Solving a time-space network formulation for the convoy movement problem. Operations research 53:219–230.
6. Beckmann M, McGuire CB, Winsten CB (1956) Studies in the Economics of Transportation (Yale University Press, New Haven).
7. Wardrop JG (1952) Some theoretical aspects of road traffic research. Proceedings of the Institute of Civil Engineers, Part II 1:325–378.
8. Bellman R (1958) On a routing problem. Q. Appl. Math. 16:87.
9. Dijkstra EW (1959) A note on two problems in connexion with graphs. Numerische Mathematik 1:269–271.
10. Xie H, Qiu L, Yang YR, Zhang Y (2004) On self adaptive routing in dynamic environments. Proc. IEEE ICNP 12-23.
11. Zhu Y, Dovrolis C, Ammar M (2006) Dynamic overlay routing based on available bandwidth estimation: A simulation study. Computer Networks Journal (Elsevier) 50:739–876.
12. Baliga J, Hinton K, Tucker RS (2007) Energy consumption of the internet. Proceedings of the Conference on Optical Internet and the 32nd Australian Conference on Optical Fibre Technology pp 1–3.
13. Shahrokhi F, Matula DW (1990) The maximum concurrent flow problem. Journal of the Association for Computing Machinery 37:318–334.
14. Leighton T, Rao S (1988) An approximate max-flow min-cut theorem for uniform multicommodity flow problems with applications to approximation algorithms. Proceedings of Foundations of Computer Science 29th Annual Symposium pp 422 – 431.
15. Awerbuch B, Azar Y, Plotkin S (1993) Throughput-competitive on-line routing. Proceedings of Foundations of Computer Science 34th Annual Symposium pp 32–40.
16. Garg N, Könemann J (1998) Faster and simpler algorithms for multicommodity flow and other fractional packing problems. Proceedings of Foundations of Computer Science 39th Annual Symposium.
17. Awerbuch B, Khandekar R (2009) Greedy distributed optimization of multi-commodity flows. Distributed Computing 317-329:317–329.
18. Barnhart C, Hane CA, Vance PH (2000) Using branch-and-price-and-cut to solve origin destination integer multicommodity flow problems. Operations Research 48:318–326.
19. Castro J, Nabona N (1996) An implementation of linear and nonlinear multicommodity network flows. European Journal of Operational Research 92:37–53.
20. Daoud M, et al. (1975) Solutions of flexible polymers. neutron experiments and interpretation. Marcomolecules 8:804.
21. Mézard M, Parisi G, Virasoro MA (1987) Spin Glass Theory and Beyond (World Scientific, Singapore).
22. Nishimori H (2001) Statistical Physics of Spin Glasses and Information Processing (Oxford University Press, Oxford ,UK).
23. Pearl J (1982) Reverend bayes on inferenced engines: A distributed hierarchical approach. Proceedings of the Second National Conference on Artificial Intelligence, Pittsburg, USA pp 133–136.
24. Gallager RG (1998) Information Theory and Reliable Communication (Wiley, NY).
25. Barash Y, et al. (2010) Deciphering the splicing code. Nature 465:53.
26. Mézard M, Parisi G, Zecchina R (2002) Analytic and algorithmic solution of random satisfiability problems. Science 297:812.
27. Liu Y, Slotine J, Barabási A (2011) Controllability of complex networks. Nature 473:167–173.
28. Bohn S, Magnasco MO (2007) Structure, scaling, and phase transition in the optimal transport network. Phys. Rev. Lett. 98:088702.
29. Banavar JR, Colaiori F, Flammani A, Maritan A, Rinaldo A (2000) Topology of the fittest transportation network. Phys. Rev. Lett 84:4745.
30. Shao Z, Zhou H (2007) Optimal transportation network with concave cost functions: loop analysis and algorithms. Phys. Rev. E 75:066112.
31. Batchelor MT, Nienhuis B, Warnaar SO (1989) Bethe-ansatz results for a solvable o(n) model on the square lattice. Phys. Rev. Lett. 62:2425–2428.
32. Stilck JF, Machado KD, Serra P (1996) Nature of the collapse transition for polymers,. Phys. Rev. Lett. 76:2734–2737.
33. Mézard M, Montanari A (2009) Information, Physics, and computation (Oxford University Press, Oxford ,UK).
34. Yeung CH, Saad D (2012) Competition for shortest paths on sparse graphs. Phys. Rev. Lett. 108:208701.
35. (Accessed in April, 2012) Transport for london, oyster card data are obtained from the website of transport for london. ("http://www.tfl.gov.uk").
36. (Accessed in April, 2012) Openflights.org, airport network data are obtained from the website of openflights.org. ("http://openflights.org/data.html").
37. Newman MEJ, Strogatz SH, Watts DJ (2001) Random graphs with arbitrary degree distributions and their applications. Phys. Rev. E 64:026118.
38. Kim M, Medard M (2004) Robustness in large-scale random networks, proceedings of the ieee infocom conference. 23-rd Annual Joint Conference of the IEEE Computer and Communications Societies 4:2364–2373.
39. Bollobás B (1985) Random Graphs (Cambridge University Press).