Papillon: Greedy Routing in Rings

Ittai Abraham* Dahlia Malkhi† Gurmeet Singh Manku‡

Abstract

We study GREEDY routing over n nodes placed in a ring, with the *distance* between two nodes defined to be the clockwise or the absolute distance between them along the ring. Such graphs arise in the context of modeling social networks and in routing networks for peer-to-peer systems. We construct the first network over n nodes in which GREEDY routing takes $O(\log n/\log d)$ hops in the worst-case, with d out-going links per node. Our result has the first asymptotically optimal greedy routing complexity. Previous constructions required $O(\frac{\log^2 n}{d})$ hops.

1 Introduction

We study GREEDY routing over uni-dimensional metrics[1] defined over n nodes lying in a ring. GREEDY routing is the strategy of forwarding a message along that out-going edge that minimizes the *distance* remaining to the destination:

> DEFINITION (Greedy Routing). *In a graph (V, E) with a given distance function $\delta : V \times V \to \mathcal{R}^+$, GREEDY routing entails the following decision: Given a target node t, a node u with neighbors $N(u)$ forwards a message to its neighbor $v \in N(u)$ such that $\delta(v,t) = \min_{x \in N(u)} \delta(x,t)$.*

Two *natural* distance metrics over n nodes placed in a circle are the clockwise-distance and the absolute-distance between pairs of nodes:

$$\begin{aligned} \delta_{clockwise}(u,v) &= \begin{cases} v-u & v \geq u \\ n+v-u & \text{otherwise} \end{cases} \\ \delta_{absolute}(u,v) &= \begin{cases} \min\{v-u, n+u-v\} & v \geq u \\ \min\{u-v, n+v-u\} & \text{otherwise} \end{cases} \end{aligned}$$

In this paper, we study the following related problems for the above distance metrics:

I. *Given integers d and Δ, what is the largest graph that satisfies two constraints: the out-degree of any node is at most d, and the length of the longest* GREEDY *route is at most Δ hops?*

II. *Given integers d and n, design a network in which each node has out-degree at most d such that the length of the longest* GREEDY *route is minimized.*

*School of Computer Science and Engineering, Hebrew University of Jerusalem, Israel. E-Mail: ittaia@cs.huji.ac.il

†Microsoft Research, Silicon Valley and School of Computer Science and Engineering, Hebrew University of Jerusalem, Israel. E-Mail: dalia@microsoft.com

‡Google Inc., USA. E-mail: manku@google.com

[1] The principles of this work can be extended to higher dimensional spaces. We focus on one-dimension for simplicity.

Summary of results

1. We construct a family of network topologies, the *Papillon*[2], in which GREEDY routes are asymptotically optimal. For both $\delta_{clockwise}$ and $\delta_{absolute}$, Papillon has GREEDY routes of length $\Delta = \Theta(\log n/\log d)$ hops in the worst-case when each node has d out-going links. Papillon is the first construction that achieves asymptotically optimal worst-case GREEDY routes.

2. Upon further investigation:, two properties of Papillon emerge: (a) GREEDY routing does not send messages along shortest paths, and (b) Edge congestion with GREEDY routing is not uniform – some edges are used more often than others. We exhibit the first property by identifying routing strategies that result in paths shorter than those achieved by GREEDY routing. In fact, one of these strategies guarantees uniform edge-congestion.

3. Finally, we consider another distance function $\delta_{xor}(u, v)$, defined as the number of bit-positions in which u and v differ. δ_{xor} occurs naturally, e.g., in hypercubes, and GREEDY routing with δ_{xor} routes along shortest paths in them. We construct a variant of Papillon that supports asymptotically optimal routes of length $\Theta(\log n/\log d)$ in the worst-case, for GREEDY routing with distance function δ_{xor}.

2 Related Work

GREEDY routing is a fundamental strategy in network theory. It enjoys numerous advantages. It is completely decentralized, in that any node takes routing decisions locally and independently. It is oblivious, thus message headers need not be written along the route. It is inherently fault tolerant, as progress toward the target is guaranteed so long as some links are available. And it has good locality behavior in that every step decreases the distance to the target. Finally, it is simple to implement, yielding robust deployments. For these reasons, GREEDY routing has long attracted attention in the research of network design. Recently, GREEDY routing has witnessed increased research interest in the context of decentralized networks. Such networks arise in modeling social networks that exhibit the "small world phenomenon", and in the design of overlay networks for peer-to-peer (P2P) systems. We now summarize known results pertaining to GREEDY routing on a circle.

The Role of the Distance Function

Efficient graph constructions are known that support GREEDY routing with distance function other than $\delta_{clockwise}$, $\delta_{absolute}$ and δ_{xor}. For de Bruijn networks, the traditional routing algorithm (which routes almost always along shortest paths) corresponds to GREEDY routing with $\delta(u, v)$ defined as the longest suffix of u that is also the prefix of v. For a 2D grid, shortest paths correspond to GREEDY routing with $\delta(u, v)$ defined as the Manhattan distance between nodes u and v.

For GREEDY routing on a circle, the best-known constructions have $d = \Theta(\log n)$ and $\Delta = \Theta(\log n)$. Examples include: Chord [SMK$^+$01] with distance-function $\delta_{clockwise}$, a variant of Chord with "bidirectional links" [GM04] and distance-function $\delta_{absolute}$, and the hypercube with distance function δ_{xor}. In this paper, we improve upon all of these constructions by showing how to route in $\Theta(\log n/\log d)$ hops in the worst case with d links per node.

GREEDY Routing in Deterministic Graphs

The Degree-Diameter Problem, studied in extremal graph theory, seeks to identify the largest graph with diameter Δ, with each node having out-degree at most d (see Delorme [D04] for a survey). The best

[2] Our constructions are variants of the well-known butterfly family, hence the name Papillon.

constructions for large Δ tend to be sophisticated [BDQ92, CG92, E01]. A well-known upper bound is $N(d, \Delta) = 1 + d + d^2 + \cdots + d^\Delta = \frac{d^{\Delta+1}-1}{d-1}$, also known as the Moore bound. A general lower bound is $d^\Delta + d^{\Delta-1}$, achieved by Kautz digraphs [K68, K69], which are slightly superior to de Bruijn graphs [dB46] whose size is only d^Δ. Thus it is possible to route in $O(\log n / \log d)$ hops in the worst-case with d out-going links per node. Whether GREEDY routes with distance functions $\delta_{clockwise}$ or $\delta_{absolute}$ can achieve the same bound, is the question we have addressed in this paper.

GREEDY routing with distance function $\delta_{absolute}$ has been studied for Chord [GM04], a popular topology for P2P networks. Chord has 2^b nodes, with out-degree $2b - 1$ per node. The longest GREEDY route takes $\lfloor b/2 \rfloor$ hops. In terms of d and Δ, the largest-sized Chord network has $n = 2^{2\Delta+1}$ nodes. Moreover, d and Δ cannot be chosen independently – they are functionally related. Both d and Δ are $\Theta(\log n)$. Analysis of GREEDY routing of Chord leaves open the following question:

For GREEDY *routing on a circle, is $\Delta = \Omega(\log n)$ when $d = O(\log n)$?*

Xu *et al.* [XKY03] provide a partial answer to the above question by studying GREEDY routing with distance function $\delta_{clockwise}$ over *uniform* graph topologies. A graph over n nodes placed in a circle is said to be uniform if the set of clockwise offsets of out-going links is identical for all nodes. Chord is an example of a uniform graph. Xu *et al.* show that for any uniform graph with $O(\log n)$ links per node, GREEDY routing with distance function $\delta_{clockwise}$ necessitates $\Omega(\log n)$ hops in the worst-case.

Cordasco *et al.* [CGH$^+$04] extend the result of Xu *et al.* [XKY03] by showing that GREEDY routing with distance function $\delta_{clockwise}$ in a uniform graph over n nodes satisfies the inequality $n \leq F(d + \Delta + 1)$, where d denotes the out-degree of each node, Δ is the length of the longest GREEDY path, and $F(k)$ denotes the k^{th} Fibonacci number. It is well-known that $F(k) = [\phi^k/\sqrt{5}]$, where $\phi = 1.618\ldots$ is the Golden ratio and $[x]$ denotes the integer closest to real number x. It follows that $1.44 \log_2 n \leq d + \Delta + 1$. Cordasco *et al.* show that the inequality is strict if $|d - \Delta| > 1$. For $|d - \Delta| \leq 1$, they construct uniform graphs based upon Fibonacci numbers which achieve an optimal tradeoff between d and Δ.

The results in [GM04, XKY03, CGH$^+$04] leave open the question whether there exists any graph construction that permits GREEDY routes of length $\Theta(\log n / \log d)$ with distance function $\delta_{clockwise}$ and/or $\delta_{absolute}$. Papillon provides an answer to the problem by constructing a non-uniform graph — the set of clockwise offsets of out-going links is different for different nodes.

GREEDY **Routing in Randomized Graphs**

GREEDY routing over nodes arranged in a ring with distance function $\delta_{clockwise}$ has recently been studied for certain classes of *randomized* graph constructions. Such graphs arise in modeling social networks that exhibit the "small world phenomenon", and in the design of overlay networks for P2P systems.

In the seminal work of Kleinberg [K00], a randomized graph was constructed in order to explain the "small world phenomenon", first identified by Milgram [M67]. The phenomenon refers to the observation that individuals are able to route letters to unknown targets on the basis of knowing only their immediate social contacts. Kleinberg considers a set of nodes on a uniform two-dimensional grid. It proposes a link model in which each node is connected to its immediate grid neighbors, and in addition, has a single long range link drawn from a normalized harmonic distribution with power 2. In the resulting graph, GREEDY routes have length at most $O(\log^2 n)$ hops in expectation; this complexity was later shown to be tight by Barrière *et al.* in [BFKK01].

Kleinberg's construction has found applications in the design of overlay routing networks for Distributed Hash Tables. Symphony [MBR03] is an adaptation of Kleinberg's construction in a single dimension. The idea is to place n nodes in a virtual circle and to equip each node with $d \geq 1$ out-going links. In the resulting network, the average path length of GREEDY routes with distance function $\delta_{clockwise}$ is $O(\frac{1}{d} \log^2 n)$ hops. Note that unlike Kleinberg's network, the space here is virtual and so are the distances and the sense of GREEDY routing. The same complexity was achieved with a slightly different

Kleinberg-style construction by Aspnes *et al.* [ADS02]. In the same paper, it was also shown that any symmetric, randomized degree-d network has $\Omega(\frac{\log^2 n}{d \log\log n})$ GREEDY routing complexity.

Papillon outperforms all of the above randomized constructions, using degree d and achieving $\Theta(\log n/\log d)$ routing. It should be possible to randomize Papillon along similar principles to the Viceroy [MNR02] randomized construction of the butterfly network, though we do not pursue this direction here.

Summary of Known Results

With $\Theta(\log n)$ out-going links per node, several graphs over n nodes in a circle support GREEDY routes with $\Theta(\log n)$ GREEDY hops. Deterministic graphs with this property include: (a) the original Chord [SMK$^+$01] topology with distance function $\delta_{clockwise}$, (b) Chord with edges treated as bidirectional [GM04] with distance function $\delta_{absolute}$. This is also the known lower bound on any uniform graph with distance function $\delta_{clockwise}$ [XKY03]. Randomized graphs with the same tradeoff include randomized-Chord [GGG$^+$03, ZGG03] and Symphony [MBR03] – both with distance function $\delta_{clockwise}$. With degree $d \le \log n$, Symphony [MBR03] has GREEDY routes of length $\Theta((\log^2 n)/d)$ on average. The network of [ADS02] also supports GREEDY routes of length $O((\log^2 n)/d)$ on average , with a gap to the known lower bound on their network of $\Omega(\frac{\log^2 n}{d \log\log n})$.

The above results are somewhat discouraging, because routing that is **non**-GREEDY can achieve much better results. In particular, networks of degree 2 with hop complexity $O(\log n)$ are well known, e.g., the Butterfly and the de Bruijn (see for example [L92] for exposition material). And networks of logarithmic degree can achieve $O(\log n/\log\log n)$ routing complexity (e.g., take the degree-$\log_2 n$ de Bruijn). Routing in these networks is non-GREEDY according to any one of our metrics ($\delta_{clockwise}$, $\delta_{absolute}$, and δ_{xor}).

The Papillon construction demonstrates that we can indeed design networks in which GREEDY routing along these metrics has asymptotically optimal routing complexity. Our contribution is a family of networks that extends the Butterfly network family, so as to facilitate efficient GREEDY routing. With d links per node, GREEDY routes are $\Theta(\log n/\log d)$ in the worst-case, which is asymptotically optimal. For $d = o(\log n)$, this beats the lower bound of [ADS02] on symmetric, randomized greedy routing networks (and it meets it for $d = O(\log n)$. In the specific case of $d = \log n$, our greedy routing achieves $O(\log n/\log\log n)$ average route length.

GREEDY **with** LOOKAHEAD

Recent work [MNW04] explores the surprising advantages of GREEDY with LOOKAHEAD in randomized graphs over n nodes in a circle. The idea behind LOOKAHEAD is to take neighbor's neighbors into account to make routing decisions. It shows that greedy with LOOKAHEAD achieves $O(\log^2 n/d \log d)$ expected route length in Symphony [MBR03]. For other networks which have $\Theta(\log n)$ out-going links per node, e.g., randomized-Chord [GGG$^+$03, ZGG03], randomized-hypercubes [GGG$^+$03], skip-graphs [AS03] and SkipNet [HJS$^+$03], average path length is $\Theta(\log n/\log\log n)$ hops. Among these networks, Symphony and randomized-Chord use GREEDY routing with distance function $\delta_{clockwise}$. Other networks use a different distance function (none of them uses δ_{xor}). For each of these networks, with $O(\log n)$ out-going links per node, it was established that plain GREEDY (*without* LOOKAHEAD) is sub-optimal and achieves $\Omega(\log n)$ expected route lengths. The results suggest that LOOKAHEAD has significant impact on GREEDY routing.

Unfortunately, realizing GREEDY routing with LOOKAHEAD on a degree-k network implies that $O(k^2)$ nodes need to be considered in each hop, while plain GREEDY needs to consider only k nodes. For $k = \log_2 n$, this implies a $O(\log n)$ overhead for LOOKAHEAD routing in every hop.

Papillon demonstrates that it is possible to construct a graph in which each node has degree d and in which GREEDY *without* 1-LOOKAHEAD has routes of length $\Theta(\log n/\log d)$ in the worst case, for the metrics $\delta_{clockwise}$, $\delta_{absolute}$ and δ_{xor}. Furthermore, for all $d = o(\log n)$, plain GREEDY on our network design beats even the results obtained in [MNW04] with 1-LOOKAHEAD.

Previous Butterfly-based Constructions

Butterfly networks have been used in the context of routing networks for DHTs as follows:

1. Deterministic butterflies have been proposed for DHT routing by Xu *et al.* [XKY03], who subsequently developed their ideas into Ulysses [KMXY03]. Papillon for distance function $\delta_{clockwise}$ has structural similarities with Ulysses – both are butterfly-based networks. The key differences are as follows: (a) Ulysses does not use $\delta_{absolute}$ as its distance function, (b) Ulysses does not use GREEDY routing, and (c) Ulysses uses more links than Papillon for distance function $\delta_{clockwise}$ – additional links have been introduced to ameliorate non-uniform edge congestion caused by Ulysses' routing algorithm. In contrast, the CONGESTION-FREE routing algorithm developed in §4 obviates the need for any additional links in Papillon (see Theorem 5).

2. Viceroy [MNR02] is a *randomized* butterfly network which routes in $O(\log n)$ hops in expectation with $\Theta(1)$ links per node. Mariposa (see reference [M04] or [M03]) improves upon Viceroy by providing routes of length $O(\log n/\log d)$ in the worst-case, with d out-going links per node. Viceroy and Mariposa are different from other randomized networks in terms of their design philosophy. The Papillon topology borrows elements of the geometric embedding of the butterfly in a circle from Viceroy [MNR02] and from [M03], while extending them for GREEDY routing.

3 Papillon

We construct two variants of butterfly networks, one each for distance-functions $\delta_{clockwise}$ and $\delta_{absolute}$. The network has n nodes arbitrarily positioned on a ring. We label the nodes from 0 to $n-1$ according to their order on the ring. For convenience, $x \bmod n$ always represents an element lying in the range $[0, n-1]$ (even when x is negative, or greater than $n-1$).

DEFINITION (Papillon for $\delta_{clockwise}$). *$\mathcal{B}_{clockwise}(\kappa, m)$ is a directed graph, defined for any pair of integers $\kappa, m \geq 1$*

1. *Let $n = \kappa^m m$.*

2. *Let $\ell(u) \equiv (m-1) - (u \bmod m)$. Each node has κ links. For node u, these directed links are to nodes $(u+x) \bmod n$, where $x \in \{1 + im\kappa^{\ell(u)} \mid i \in [0, \kappa-1]\}$.*

   We denote the link with node $(u+1) \bmod n$ as u's "short link"*. The other $\kappa - 1$ links are called u's* "long links"*.*

DEFINITION (Papillon for $\delta_{absolute}$). *$\mathcal{B}_{absolute}(k, m)$ is a directed graph, defined for any pair of integers $k, m \geq 1$,*

1. *Let $n = (2k+1)^m m$.*

2. *Let $\ell(u) \equiv (m-1) - (u \bmod m)$. Each node has $2k+2$ out-going links. Node u makes $2k+1$ links with nodes $(u+x) \bmod n$, where $x \in \{1 + im(2k+1)^{\ell(u)} \mid i \in [-k, +k]\}$. Node u also makes an out-going link with node $(u+x) \bmod n$, where $x = -m+1$.*

   We denote the link with node $(u+1) \bmod n$ as u's "short link"*. The other $2k+1$ links are called u's* "long links"*.*

In both $\mathcal{B}_{clockwise}$ and $\mathcal{B}_{absolute}$, all out-going links of node u are incident upon nodes with level $(\ell(u)-1) \bmod m$. In $\mathcal{B}_{clockwise}$, the short links are such that each hop diminishes the remaining *clockwise* distance by at least one. Therefore, GREEDY routing is guaranteed to take a finite number of hops. In $\mathcal{B}_{absolute}$, not every GREEDY hop diminishes the remaining *absolute* distance. However, GREEDY routes are still finite in length, as we show in the proof of Theorem 2.

Theorem 1. GREEDY *routing in* $\mathcal{B}_{clockwise}$ *with distance function* $\delta_{clockwise}$ *takes* $3m-2$ *hops in the worst-case. The average is less than* $2m-1$ *hops.*

Proof. For any node u, we define $\text{SPAN}(u) \equiv \{v \mid 0 \leq \delta_{clockwise}(u,v) < m\kappa^{\ell(u)+1}\}$. Let t and u denote the target node and the current node, respectively. Routing proceeds in (at most) three phases:

Phase I:	$t \notin \text{SPAN}(u)$	(at most $m-1$ hops)
Phase II:	$t \in \text{SPAN}(u)$ and $\delta_{clockwise}(u,t) \geq m$	(at most m hops)
Phase III:	$t \in \text{SPAN}(u)$ and $\delta_{clockwise}(u,t) < m$	(at most $m-1$ hops)

We now prove upper bounds on the number of hops in each phase.

I. The out-going links of u are incident upon nodes at level $(\ell(u)-1) \bmod m$. So eventually, the level of the current node u will be $m-1$. At this point, $t \in \text{SPAN}(u)$ because $\text{SPAN}(u)$ includes *all* the nodes. Thus Phase 1 lasts for at most $m-1$ hops ($\frac{m-1}{2}$ hops on average).

II. GREEDY will forward the message to some node v such that $t \in \text{SPAN}(v)$ and $\ell(v) = \ell(u)-1$. Eventually, the current node u will satisfy the property $\ell(u) = 0$. This node will forward the message to some node v with $\ell(v) = m-1$ such that $\delta_{clockwise}(v,t) < m$, thereby terminating this phase of routing. There are at most m hops in this phase (at most m on average as well).

III. In this phase, GREEDY will decrease the clockwise distance by exactly one in each hop by following the short-links. Eventually, target t will be reached. This phase takes at most $m-1$ hops ($\frac{m-1}{2}$ hops on average).

The worst-case route length is $3m-2$. On average, routes are at most $2m-1$ hops long. □

Theorem 2. GREEDY *routing in* $\mathcal{B}_{absolute}$ *with distance function* $\delta_{absolute}$ *takes* $3m-2$ *hops in the worst-case. The average is less than* $2m-1$ *hops.*

Proof. For any node u, we define

$$\text{SPAN}(u) \equiv \{v \mid \delta_{absolute}(u,v) = |c + m\sum_{i=0}^{\ell(u)}(2k+1)^i d_i|,\ c \in [0, m-1],\ d_i \in [-k, +k]\}.$$

Let t and u denote the target node and the current node, respectively.
Routing proceeds in (at most) three phases:

Phase I:	$t \notin \text{SPAN}(u)$	(at most $m-1$ hops)
Phase II:	$t \in \text{SPAN}(u)$ and $\delta_{absolute}(u,t) \geq m$	(at most m hops)
Phase III:	$t \in \text{SPAN}(u)$ and $\delta_{absolute}(u,t) < m$	(at most $m-1$ hops)

We now prove upper bounds on the number of hops in each phase.

I. All out-going links of node u are incident upon nodes at level $(\ell(u)-1) \bmod m$. So eventually, the current node u will satisfy the property $\ell(u) = m-1$. At this point, $t \in \text{SPAN}(u)$ because $\text{SPAN}(u)$ includes *all* nodes. Thus Phase I lasts at most $m-1$ hops (at most $\frac{m-1}{2}$ hops on average).

II. Phase 2 terminates if target node t is reached, or if $\delta_{absolute}(u,t) < m$. Node u always forwards the message to some node v such that $t \in \text{SPAN}(v)$ and $\ell(v) = \ell(u) - 1$. So eventually, either target t is reached, or the current node u satisfies the property $\ell(u) = 0$. At this point, if node u forwards the message to node v, then it is guaranteed that $\ell(v) = m - 1$ and $\delta_{absolute}(v,t) < m$, thereby terminating Phase II. There are at most m hops in this phase (at most m on average as well).

III. The target node t is reached in at most $m-1$ hops (the existence of the "back edge" that connects node u to node $(u+1-m) \bmod n$ guarantees this). This phase takes at most $m-1$ hops (at most $\frac{m-1}{2}$ hops on average).

The worst-case route length is $3m-2$. On average, routes are at most $2m-1$ hops long. □

Routes in both $\mathcal{B}_{clockwise}$ and $\mathcal{B}_{absolute}$ are at most $3m-2$ hops, which is $O(\log(\kappa^m m)/\log\kappa)$ and $O(\log((2k+1)^m m)/\log(2k+2))$, respectively. Given degree d and diameter Δ, the size of Papillon is $n = 2^{O(\Delta)}\Delta$ nodes. Given degree d and network size n, the longest route has length $\Delta = O(\log n/\log d)$.

4 Improved Routing Algorithms for Papillon

GREEDY routing does not route along shortest-paths in $\mathcal{B}_{clockwise}$ and $\mathcal{B}_{absolute}$. We demonstrate this constructively below, where we study a routing strategy called HYPERCUBIC-ROUTING which achieves shorter path lengths than GREEDY.

Hypercubic Routing

Theorem 3. *There exists a routing strategy for $\mathcal{B}_{clockwise}$ in which routes take $2m-1$ hops in the worst-case. The average is at most $1.5m$ hops.*

Proof. Consider the following HYPERCUBIC-ROUTING algorithm on $\mathcal{B}_{clockwise}$. Let s be the source node, t the target, and let $dist = \delta_{clockwise}(s,t) = c + m + m\sum_{i=0}^{i=m-1}\kappa^i d_i$ with $0 \le c < m$ and $0 \le d_i < \kappa$ ($dist$ has exactly one such representation, unless $dist \le m$ in which case routing takes $< m$ hops).

Phase I: Follow the short-links to "fix" the c-value to zero. This takes at most $m-1$ hops (at most $0.5m$ hops on average).

Phase II: In exactly m hops, "fix" the d_i's in succession to make them all zeros: When the current node is u, we fix $d_{\ell(u)}$ to zero by following the appropriate long-link, i.e., by shrinking the clockwise distance by $d_{\ell(u)}\kappa^{\ell(u)}m + 1$. The new node v satisfies $\ell(v) = (\ell(u) + m - 1)(\text{mod } m)$. When each d_i is zero, we have reached the target.

Overall, the worst-case route length is $2m-1$. Average route length is at most $1.5m$. □

Theorem 4. *There exists a routing strategy for $\mathcal{B}_{absolute}$ in which routes take $2m-1$ hops in the worst-case. The average is at most $1.5m$ hops.*

Proof. Let s be the source node, t the target.

Phase I: Follow the short-links in the clockwise direction, to reach a node s' such that $\ell(s') = \ell(t)$. This takes at most $m-1$ hops (at most $0.5m$ hops on average). The remaining distance can be expressed as $m + m\sum_{i=0}^{i=m-1}(2k+1)^i d_i$ where $-k \le d_i \le k$. There is a unique such representation.

Phase II: In exactly m hops, "fix" the d_i's in succession to make them all zeros: When the current node is u, we fix $d_{\ell(u)}$ by following the appropriate long-link, i.e., by traveling distance $1 + d_{\ell(u)}(2k+1)^{\ell(u)}m$ along the circle (this distance is positive or negative, depending upon the sign of $d_{\ell(u)}$). The new node v satisfies $\ell(v) = (\ell(u) - 1)(\text{mod } m)$. When each d_i is zero, we have reached the target.

Overall, the worst-case route length is $2m-1$. Average route length is at most $1.5m$. □

Note that the edges that connect node u to node $(u+1-m) \bmod n$ are redundant for HYPERCUBIC-ROUTING since they are never used. However, these edges play a crucial role in GREEDY routing in $\mathcal{B}_{absolute}$ (to guide the message to the target in Phase 3).

Congestion-Free Routing

Theorems 3 and 4 prove that GREEDY routing is sub-optimal in the constants. HYPERCUBIC-ROUTING, as described above, is faster than GREEDY. However, it causes *edge-congestion* because short-links are used more often than long-links. Let π denote the ratio of maximum and minimum loads on edges caused by all $\binom{n}{2}$ pairwise routes. HYPERCUBIC-ROUTING for $\mathcal{B}_{clockwise}$ consists of two phases (see Proof of Theorem 3). The load due to Phase II is uniform – all edges (both short-links and long-links) are used equally. However, Phase I uses only short-links, due to which $\pi \neq 1$. We now modify the routing scheme slightly to obtain $\pi = 1$ for both $\mathcal{B}_{clockwise}$ and $\mathcal{B}_{absolute}$.

Theorem 5. *There exists a congestion-free routing strategy in $\mathcal{B}_{clockwise}$ that takes $2m-1$ hops in the worst-case and at most $1.5m$ hops on average, in which $\pi = 1$.*

Proof. The theorem is proved constructively, by building a new routing strategy called CONGESTION-FREE. This routing strategy is exactly the same as HYPERCUBIC-ROUTING, with a small change.

Let s be the source node, t the target. Let $c = (t+m-s) \bmod m$, the difference in levels between $\ell(s)$ and $\ell(t)$.

Phase I: For c steps, follow any out-going link, chosen uniformly at random. We thus reach a node s' such that $\ell(s') = \ell(t)$.

Phase II: The remaining distance is $dist = \delta_{clockwise}(s',t) = m + m\sum_{i=0}^{i=m-1} \kappa^i d_i$ with $0 \leq d_i < \kappa$. Continue with Phase II of the HYPERCUBIC-ROUTING algorithm for $\mathcal{B}_{clockwise}$ (see Theorem 3).

It is easy to see that in this case, all outgoing links (short- and long-) are used with equal probability along the route. Hence, $\pi = 1$. □

Theorem 6. *There exists a congestion-free routing strategy in $\mathcal{B}_{absolute}$ that takes $2m-1$ hops in the worst-case and at most $1.5m$ hops on average, in which $\pi = 1$.*

Proof. We will ignore the edges that connect node u to node $(u+1-m) \bmod n$ (recall that these edges are not used in HYPERCUBIC-ROUTING described in Theorem 4). We will ensure $\pi = 1$ for the remainder of the edges.

CONGESTION-FREE routing follows the same idea as that for $\mathcal{B}_{clockwise}$ (Theorem 5): Let s be the source node, t the target. Let $c = (t+m-s) \bmod m$, the difference in levels between $\ell(s)$ and $\ell(t)$. In Phase I, for c steps, we follow any out-going link, chosen uniformly at random. We thus reach a node s' such that $\ell(s') = \ell(t)$. In Phase II, we continue as per Phase II of the HYPERCUBIC-ROUTING algorithm for $\mathcal{B}_{absolute}$ (Theorem 4).

An alternate CONGESTION-FREE routing algorithm for $\mathcal{B}_{absolute}$ that routes deterministically is based upon the following idea: We express any integer $a \in [-k,+k]$ as the sum of two integers: $a' = \lfloor (k+a)/2 \rfloor$ and $a'' = -\lfloor (k-a)/2 \rfloor$. It is easy to verify that $a = a' + a''$. Now if we list all pairs $\langle a', a'' \rangle$ for $a \in [-k,+k]$, then each integer in the range $[-k,+k]$ appears exactly twice as a member of some pair.

Let s be the source node, t the target. Let $c = (t+m-s) \bmod m$, the difference in levels between $\ell(s)$ and $\ell(t)$. The remaining distance is $dist = c + m + m\sum_{i=0}^{i=m-1}(2k+1)^i d_i$ with $-k \leq d_i \leq k$ (there is a unique way to represent $dist$ in this fashion).

Phase I: For c steps, if the current node is u, then we follow the edge corresponding to $d'_{\ell(u)}$, i.e., the edge that covers distance $1 + md'_{\ell(u)}(2k+1)^{\ell(u)}$ (in the clockwise or the anti-clockwise direction, depending upon the sign of $d'_{\ell(u)}$). At the end of this phase, we reach a node s' such that $\ell(s') = \ell(t)$.

Phase II: Continue with Phase II of the HYPERCUBIC-ROUTING algorithm for $\mathcal{B}_{absolute}$ (Theorem 4), for exactly m steps.

Due to the decomposition of integers in $[-k, +k]$ into pairs, as defined above, all outgoing links (short- and long-) are used equally. Hence, $\pi = 1$. □

Notes: In the context of the current Internet, out-going links correspond to full-duplex TCP connections. Therefore, the undirected graph corresponding to $\mathcal{B}_{absolute}$ is of interest. In this undirected graph, it is possible to devise congestion-free routing with $\pi = 1$, maximum path length $m + \lfloor m/2 \rfloor$ and average route-length at most $1.25m$. This is achieved by making at most $\lfloor m/2 \rfloor$ initial random steps either in the down or the up direction, whichever gets to a node with level $\ell(t)$ faster.

5 Papillon with Distance Function δ_{xor}

In this Section, we define a variant of Papillon in which GREEDY routing with distance function δ_{xor} results in worst-case route length $\Theta(\log n / \log d)$, with n nodes, each having d out-going links. For integers s and t, $\delta_{xor}(s, t)$ is defined as the number of bit-positions in which the binary representations of s and t differ.

DEFINITION (Papillon for δ_{xor}). *$\mathcal{B}_{xor}(\lambda, m)$ is a directed graph, defined for any pair of integers $\lambda, m \geq 1$ where λ is a power of two.*

1. *The network has $n = m\lambda^m$ nodes labeled from 0 to $n - 1$.*
2. *Let u denote a node. Let $\ell(u)$ denote the unique integer $x \in [0, m-1]$ that satisfies $x\lambda^m \leq u < (x+1)\lambda^m$. The node u makes links with nodes with labels*

$$((\ell(u) + 1) \bmod m)\lambda^m + i\lambda^{\ell(u)}, \quad \text{where } \; 0 \leq i < \lambda.$$

Thus, if (u, v) is an edge, then $\ell(v) = (\ell(u) + 1) \bmod m$.

Theorem 7. GREEDY *routing in $\mathcal{B}_{xor}$ with distance function δ_{xor} takes $2m - 1$ hops in the worst-case. The average is at most $1.5m$ hops.*

Proof. Let the current node be s. Let t denote the target node. Then $s \oplus t$, the bit-wise exclusive-OR of s and t, can uniquely be expressed as $c + \sum_{i=0}^{i=m-1} \lambda^i d_i$, where $c \geq 0$ and $0 \leq d_i < \lambda$. Routing proceeds in two phases. In Phase I, each of the d_i is set to zero. This takes at most m steps (at most m on average). In Phase II, the most significant $\lceil \log_2 m \rceil$ bits of $s \oplus t$ are set to zero, thereby reaching the target. This phase takes at most $m - 1$ hops (at most $\frac{m-1}{2}$ on average). □

6 Summary

We presented Papillon, a variant of multi-butterfly networks which supports asymptotically optimal GREEDY routes of length $O(\log n / \log d)$ with distance functions $\delta_{clockwise}$, $\delta_{absolute}$ and δ_{xor}, when each node makes d out-going links, in an n-node network. Papillon is the first construction with this property.

Some questions that remain unanswered:

1. *Is it possible to devise graphs in which* GREEDY *routes with distance function $\delta_{clockwise}$ and $\delta_{absolute}$ are along shortest-paths?* As Theorems 3 and 4 illustrate, GREEDY routing on Papillon do not route along shortest-paths. Is this property inherent in GREEDY routes?

2. *What is the upper-bound for the Problem of Greedy Routing on the Circle?* Papillon furnishes a lower-bound, which is asymptotically optimal. However, constructing the largest-possible graph with degree d and diameter Δ, is still an interesting combinatorial problem.

References

[ADS02] J Aspnes, Z Diamadi, and G Shah. Fault-tolerant routing in peer-to-peer systems. *Proc. 21st ACM Symposium on Principles of Distributed Computing (PODC 2002)*, p. 223–232, 2002.

[AS03] J Aspnes and G Shah. Skip graphs. *Proc. 14th ACM-SIAM Symposium on Discrete Algorithms (SODA 2003)*, p. 384–393, 2003.

[BDQ92] J C Bermond, C Delorme, and J J Quisquater. Table of large (δ, d)-graphs. *Discrete Applied Mathematics*, 37/38:575–577, 1992.

[BFKK01] L Barrière, P Fraigniaud, E Kranakis, and D Krizanc. Efficient routing in networks with long range contacts. *Proc. 15th Intl. Symposium on Distributed Computing (DISC 2001)*, p. 270–284, 2001.

[CG92] F Comellas and J Gómez. New large graphs with given degree and diameter. *Graph Theory, Combinatorics and Algorithms*, 1:221–233, 1992.

[CGH$^+$04] G Cordasco, L Gargano, M Hammar, A Negro, and V Scarano. F-Chord: Improved uniform routing on Chord. *Proc. 11th Colloquium on Structural Information and Communication Complexity*, 2004.

[D04] C Delorme. The (Degree, Diameter) problem for graphs. Laboratoire de Recherche en Informatique, Université Paris Sud, France. Available as http://maite71.upc.es/grup_de_grafs/table_g.html, 2004.

[dB46] N G d Bruijn. A combinatorial problem. *Proc. Koninklijke Nederlandse Akademie van Wetenschappen*, 49:758–764, 1946.

[E01] G Exoo. A family of graphs and the degree/diameter problem. *J. of Graph Theory*, 37:118–124, 2001.

[GGG$^+$03] K P Gummadi, R Gummadi, S D Gribble, S Ratnasamy, S Shenker, and I Stoica. The impact of DHT routing geometry on resilience and proximity. *Proc. ACM SIGCOMM 2003*, p. 381–394, 2003.

[GM04] P Ganesan and G S Manku. Optimal routing in Chord. *Proc. 15th ACM-SIAM Symposium on Discrete Algorithms (SODA 2004)*, p. 169–178, 2004.

[HJS$^+$03] N J A Harvey, M Jones, S Saroiu, M Theimer, and A Wolman. SkipNet: A scalable overlay network with practical locality properties. *Proc. 4th USENIX Symposium on Internet Technologies and Systems (USITS 2003)*, 2003.

[K68] W H Kautz. Bounds on directed (d, k) graphs. *Theory of Cellular Logic Networks and Machines (AFCRL-68-0668, SRI Project 7258, Final Report)*, p. 20–28, 1968.

[K69] W H Kautz. Design of optimal interconnection networks for multiprocessors. *Architecture and Design of Digital Computers (Nato Advanced Summer Institute)*, p. 249–272, 1969.

[K00] J Kleinberg. The small-world phenomenon: An algorithmic perspective. *Proc. 32nd ACM Symposium on Theory of Computing (STOC 2000)*, p. 163–170, 2000.

[KMXY03] A Kumar, S Merugu, J J Xu, and X Yu. Ulysses: A robust, low-diameter, low-latency peer-to-peer network. *Proc. 11th IEEE International Conference on Network Protocols (ICNP 2003)*, 2003.

[L92] F T Leighton. *Introduction to Parallel Algorithms and Architectures: Arrays - Trees - Hypercubes.* Academic Press/Morgan Kaufmann, 1992.

[M67] S Milgram. The small world problem. *Psychology Today*, 67(1):60–67, 1967.

[M03] G S Manku. Routing networks for distributed hash tables. *Proc. 22nd ACM Symposium on Principles of Distributed Computing (PODC 2003)*, p. 133–142, 2003.

[M04] G S Manku. *Dipsea: A Modular Distributed Hash Table.* PhD dissertation, Stanford University, Department of Computer Science, 2004.

[MBR03] G S Manku, M Bawa, and P Raghavan. Symphony: Distributed hashing in a small world. *Proc. 4th USENIX Symposium on Internet Technologies and Systems (USITS 2003)*, p. 127–140, 2003.

[MNR02] D Malkhi, M Naor, and D Ratajczak. Viceroy: A scalable and dynamic emulation of the butterfly. *Proc 21st ACM Symposium on Principles of Distributed Computing (PODC 2002)*, p. 183–192, 2002.

[MNW04] G S Manku, M Naor, and U Wieder. Know thy neighbor's neighbor: The power of lookahead in randomized P2P networks. *Proc. 36th ACM Symposium on Theory of Computing (STOC 2004)*, p. 54–63, 2004.

[SMK+01] I Stoica, R Morris, D Karger, M F Kaashoek, and H Balakrishnan. Chord: A scalable peer-to-peer lookup service for Internet applications. *Proc. ACM SIGCOMM 2001*, p. 149–160, 2001.

[XKY03] J Xu, A Kumar, and X Yu. On the fundamental tradeoff between routing table size and network diameter in peer-to-peer networks. *Proc. IEEE INFOCOM 2003*, 2003.

[ZGG03] H Zhang, A Goel, and R Govindan. Incrementally improving lookup latency in distributed hash table systems. *ACM SIGMETRICS 2003*, p. 114–125, 2003.