Routing at Large Scale: Advances and Challenges for Complex Networks

Sahel Sahhaf, Wouter Tavernier, Dimitri Papadimitriou, Davide Careglio, Alok Kumar, Christian Glacet, David Coudert, Nicolas Nisse, Lluís Fàbrega, Pere Vilà, Miguel Camelo, Pieter Audenaert, Didier Colle, and Piet Demeester

Abstract

A wide range of social, technological and communication systems can be described as complex networks. Scale-free networks are one of the well known classes of complex networks in which nodes' degrees follow a power-law distribution. The design of scalable, adaptive and resilient routing schemes in such networks is very challenging. In this article we present an overview of required routing functionality, categorize the potential design dimensions of routing protocols among existing routing schemes, and analyze experimental results and analytical studies performed so far to identify the main trends/trade-offs and draw main conclusions. Besides traditional schemes such as hierarchical/shortest-path path-vector routing, the article pays attention to advances in compact routing and geometric routing since they are known to significantly improve scalability in terms of memory space. The identified trade-offs and the outcomes of this overview enable more careful conclusions regarding the (un-)suitability of different routing schemes to large-scale complex networks and provide a guideline for future routing research.

Introduction

Complex networks refer to large, dynamic networks consisting of potentially billions of nodes and links that are used to describe a wide range of social, biological, technological and communication systems. Scale-free networks as one well known/much studied class of complex networks have degree distribution[1] that follows a power-law function. In such networks, new nodes attach preferentially to already well-connected nodes. The network of autonomous systems[2] (ASes) forming the core of the Internet graph, is an example of such networks.[3] Routing[4] in these networks is challenging because of the size of the network, and the properties and performance expected from these networks, particularly, any-to-any connectivity, availability, and reliability.

Routing research has evolved very pragmatically in communication networks from small scale to larger scale in technologies including wireless, ad-hoc/sensor networks, the Internet, and so on. Since new networks of increasing scale are popping up every day (e.g., Internet of Things), it is important to consider clean-slate approaches considering the entire design space of routing paradigms to avoid getting "trapped" again in legacy protocols/paradigms.

In this article we try to open this design question by clearly and cautiously categorizing/grouping the potential design dimensions of routing protocols among existing routing schemes (traditional schemes as well as novel ones), analyzing experimental results performed so far, and drawing some main conclusions, guidelines and open challenges for routing schemes in future settings.

This article synthesizes the fundamental aspects of routing schemes for complex networks, as well as lessons learned from experimental routing research stemming from the EULER project (http://www.euler-fire-project.eu/). Particular attention will be given to:

- New classes of path-based routing schemes[5].
- New routing paradigms subdivided into locator space dependent[6] and locator space independent.
- New route discovery schemes in which networks' structural properties are used.

Additionally, a brief overview of improvements to path-vector schemes and routing advances in delay/disruption-tolerant networks (DTN) and peer-to-peer overlays is provided.

This work presents an overview of the design dimensions of routing protocols, challenges, and a perspective/guideline for future routing research in complex networks.

Routing Design Problem

Routing is the process of finding/selecting paths between given nodes of a communication network. A path is a finite sequence of nodes from a source toward a destination node. The distance between two nodes is the sum of the cost of the links used along its shortest interconnecting path. The diameter of a network ($D(G)$) is the maximum distance[7] of any two nodes.

Sahel Sahhaf, Wouter Tavernier, Pieter Audenaert, Didier Colle, and Piet Demeester are with Ghent University-IMEC.

Dimitri Papadimitriou is with Nokia Bell Labs.

Davide Careglio is with Universitat Politècnica de Catalunya.

A Kumar is with Université Catholique de Louvain.

Christian Glacet, David Coudert, and Nicolas Nisse are with the Université Côte d'Azur - Inria - CNRS - I3S.

Lluís Fàbrega and Pere Vilà are with the University of Girona.

Miguel Camelo is with the University of Antwerp - IMEC.

Digital Object Identifier:
10.1109/MNET.2017.1600203

[1] The probability that a node selected uniformly at random has a certain number of links.

[2] In the Internet, an autonomous system is a single network or a group of networks that is managed and supervised by a single administrative entity or organization.

[3] The power-law component of the Internet seems to be decreasing while the assortativity (likelihood of nodes with the same degree being connected) is increasing.

[4] The process of finding/selecting paths between given nodes of a communication network.

[5] Schemes that maintain the path information to reach a destination.

[6] A routing paradigm that derives paths based on locators. Particularly it enables routing the packets based on addresses that are specific to the network location instead of relying on any arbitrary flat address space.

[7] The distance between two nodes is the number of edges in the shortest path between the nodes.

The Routing Function

Routing is decomposed into the following functionalities.

Identification and Location: In order to derive paths between nodes of a topology, nodes should be identified. A node identifier might be a number/label. Identification functionality does not necessarily imply a location within the topology.[8] Thus, the routing should focus on structuring the topological space into addresses/locators, and mapping the identifier to the network nodes locator when needed.

Discovery/Distribution: This is required to discover/distribute information related to routes or topology characteristics. It can be push-based (local changes are distributed toward remote nodes) or pull-based (on demand search) or a combination.

Policy: Policing routing, including routing-engineering, traffic-engineering and administrative policies, affects both local processing performance and the overall performance resulting from local decisions. Limiting policing capabilities leads to an increase in local performance but may decrease global performance, while increasing flexibility may increase global performance. From the routing design perspective, this leads to a major consequence: starting from a relatively simple routing procedure and requiring homogeneous policy strategy (which is unlikely in organically controlled organizations such as the Internet) may lead to detrimental effects in terms of performance.

Route Determination/Calculation: This functionality determines routes toward a given destination. This operation can involve routing path calculations constrained by policies and/or route selection/filtering functionality, or can be guided by a substructure of the discovered topology (e.g., network spanning tree).

Routing Entry Determination: This determines routing entries in the routing tables (RT)[9] based on the outcome of route determination functionality. The outcome can be a selected set of potential next hops for given network locations, or a procedure to decide how such a routing entry can be determined on the fly.

Multicast: The previous functionalities primarily target unicast routing. Multicast routing is a distributed algorithm that allows any node to route multicast traffic to a group of destination nodes, called a multicast group. To enable point-to-multipoint traffic distribution, the multicast routing protocol builds a tree between the source and the multicast group called a Multicast Distribution Tree (MDT). Multicast routing is (re-)gaining interest given the increasing popularity of multimedia streaming/content traffic, since it yields bandwidth savings competing with/complementing cached content distribution techniques. Multicast tree membership management handles multicast membership, which involves the join/subscription and leave/un-subscription actions.

[8] Note that traditional IP addresses fulfill both roles, leading to significant issues regarding node mobility, multi-homing, and so on, as confirmed by the invention of, for example, the ILNP, LISP and HIP protocols. The impact of locator/ID separation is detailed in [1].

[9] A routing table is a local data structure stored in a router. It may store routes to network destinations, neighbor node identifiers, output port numbers or other. It is used by the routing algorithm to perform routing.

A node identifier might be a number/label. Identification functionality does not necessarily imply a location within the topology.8 Thus, the routing should focus on structuring the topological space into addresses/locators, and mapping the identifier to the network nodes locator when needed.

A routing function determines the next-hop along a path from a source toward a destination. This path is determined by the routing schemes, which are described according to the following key-properties.

Uncoordinated vs. Coordinated Routing Decision: In an uncoordinated routing each node takes its routing decision independently of others though all the participating entities adhere and work toward global shared objectives, such as connectivity and availability.

Distributed vs. Centralized: Unlike a centralized algorithm, a distributed algorithm is executed locally at each node and independently of other nodes. They are different from uncoordinated algorithms as distribution is about computation while the latter refers to routing decision.

Control vs. Data-Driven: Control-driven algorithms are triggered by independent processes exchanging control information, while data-driven algorithms only trigger routing algorithms when data packets travel through the network.

Deterministic vs. Statistical: In deterministic routing, the path determination between a set of nodes is fixed and independent of time or particular data within control/data traffic between nodes. Statistical algorithms introduce a degree of randomness within the generated routes.

Stateful vs. Stateless: Unlike stateless algorithms, stateful algorithms require the maintenance of states to operate, for example, for storing information related to the interaction with other nodes.

We mainly focus on the advances to control-driven, stateful distributed routing (meaning routing information is exchanged via dedicated messages, nodes store RT entries and their computation is distributed), the other dimensions being dependent on the schemes.

Trade-offs in Routing

When routing at large-scale (above 10000 nodes), the following three cost dimensions can be identified.

Memory Cost: The memory space in a node required to store the routing information used by the routing algorithm (input) and to store the RTs (output).

Stretch Cost: Stretch is the ratio between the length of a path generated by the scheme and the corresponding shortest paths. The stretch of a routing scheme is the highest stretch among all source-destination pairs.

Adaptation Cost: Communication complexity that refers to the number of exchanged messages between nodes for the computation of the RT entries, and convergence time as the difference in time between the sending of the first message and the reception of the last message during the execution of the routing algorithm.

Upon designing a routing scheme, a trade-off should be taken into account between different criteria depicted in Fig. 1. When designing a "stat-

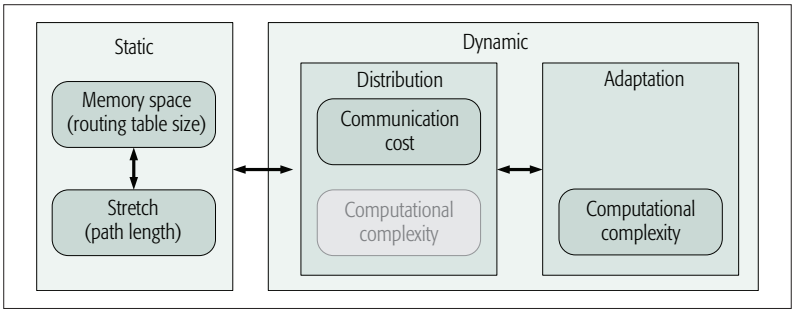

FIGURE 1. Fundamental trade-offs in routing schemes.

ic" scheme, there is a trade-off between memory space and stretch. This means that it is not possible to have 1-stretch/fully shortest-path routing, and good (sub-linear) scaling of memory at all nodes. When distributing the scheme, communication cost becomes an additional criterion impacting the previous trade-off, and moving to an adaptive scheme, computational complexity adds to it. Figure 1 also shows that when designing a "dynamic" scheme, both distribution and adaptation should be considered. Computational complexity is not the main criterion when moving to a distributed scheme (indicated with gray color).

Challenges in the Internet Routing System

Since we target large-scale complex networks, in this section we explain the main open challenges in the Internet routing as it is one known large-scale scale-free network in nature/technical domain.

The current Internet routing follows a two-level hierarchy: routing between almost 60 K ASes in the core (forming a scale-free network), and routing within the ASes. The true challenge is in the inter-AS routing, driven by the Border Gateway Protocol (BGP), which is a path-vector routing protocol,[10] exchanging network reachability information with peering BGP routers. Reachability information includes an AS path listing the sequence of AS numbers traversed by the BGP route advertisement comprising reachability information from the originating AS. Discovered path information is used by BGP routers for constructing the AS connectivity graph for this reachability, and to detect/avoid routing loops by performing a route selection process combined with shortest path routing. Policies are determined to maintain business relationships between peering ASes or by load balancing strategies during high-traffic periods.

BGP is subject to the path exploration/hunting phenomenon: BGP routers may announce as valid, routes that are affected by a failure that are withdrawn later during subsequent routing updates. This is (one of) the main reasons for the large number of update messages received by inter-domain routers.

Internet routing is significantly challenged by the increasing number of routers, ASes, and routes. This situation is exacerbated due to site multi-homing,[11] AS multi-homing, traffic engineering, and the increasing need for connectivity availability from the increasing number of connected hosts.

The main issues in the Internet architecture are the scalability, convergence, and stability properties of its inter-domain routing. Solving them requires addressing multiple dimensions altogether, for example, the RT size growth resulting from a large number of message exchanges induced by topological/policy changes. Both dimensions increase memory and processing requirements of routing engines. Solving the scalability of the Internet routing, considering its dynamics, is challenging. Convergence time should not be delayed, whereas scalability improvement minimizes the number of exchanged messages, thereby avoiding overloading the routing calculation engine.

However, when considering the existing Internet routing and the considerable research efforts made to improve it, one might wonder, when considering a complex network with similar conditions/constraints as the Internet, whether the only feasible solution is a local policy-based path-vector routing system? Or would there be a more promising model beyond routing IP packets?

Routing Schemes

We consider the following classification: improvements to path-vector schemes, routing schemes (clean-slate approaches) in complex networks, and routing schemes in DTN and P2P networks.

Improvements to Path-Vector Schemes

Numerous enhancements to path-vector schemes have been proposed over the last 20 years. BGP is an example of a path-vector protocol driving the inter-AS routing in the Internet. Many of the enhancements relate to BGP dynamic properties. Examples include:

- Enhanced Path Vector Routing protocol (EPIC), which annotates the AS paths with additional "path dependency" information to reduce convergence time.
- BGP with root cause notification reduces the convergence time by announcing the root cause of a link failure location.
- Path exploration damping augments BGP for selectively damping the propagation of path exploration updates.

Recently, new route selection schemes have been proposed to improve route stability in BGP [2].

Routing Schemes in Complex Networks

Table 1 positions our proposed routing schemes[12] (clean-slate approaches) with respect to their adaptation capability to topology/policy dynamics, distribution of the computation/decision process, communication type and addressing scheme. BGP is included for comparison. As mentioned earlier, these are control-driven, stateful and distributed routing schemes.

Compact Routing: The goal of compact routing is to reduce the amount of storage space in each node. It is challenging to design algorithms with a good trade-off between the memory space

[10] A routing protocol that maintains path information to reach the destinations. This information is updated dynamically. Using this scheme, the routing tables include the destination network, the next router and the path to the destination. It is easy to detect routing loops and discard the update messages that are looping in the network.

[11] The practice of connecting a host or a network to more than one network.

[12] Routing schemes proposed within the FP7 EULER project.

	Centralized	Distributed	Static	Fault–tolerant/adaptive	Multicast	Unicast	Locator–independent
Route discovery (RD)	–	✓	–	✓	–	✓	–
Centralized compact routing (AGMaNT)	✓	–	✓	–	–	✓	✓
Distributed compact routing (DCR) [3]	–	✓	✓	–	–	✓	✓
Greedy compact multicast routing (GCMR) [4]	–	✓	–	✓	✓	–	✓
Geometric tree-based greedy routing (GTR) [5]	–	✓	–	✓	–	✓	–
Word-metric-based greedy routing (WMGR) [6]	–	✓	–	✓	–	✓	–
Geometric coordinate-labeling scheme (GCLS) [7]	–	✓	–	✓	–	✓	–
Border gateway protocol (BGP)	–	✓	–	✓	–	✓	–

TABLE 1. Position of different routing schemes with respect to the capability to adapt to dynamics, distribution and their communication type. The last column indicates whether the scheme is locator-independent or not.

and the resulting stretch. The theoretical bounds concern worst-case analysis, and one of the contribution of EULER is to show that much better trade-offs are achieved in actual networks. We have investigated two research directions: unicast and multicast compact routing.[13]

Route Discovery with the Network's Structural Properties (RD): We designed a route discovery scheme for an inter-AS network where each network is a member of a specific group. The country code (ISO 3166) is used for defining groups in the Internet and assumed that at least one path exists between each pair of nodes. This scheme is based on limited network information, that is, two-hop neighborhood information, and membership of nodes to groups, whose efficiency is based on the existence of highly popular nodes and the similarity of adjacent nodes.

The route discovery scheme is initiated by the source node that issues a discovery packet, which is forwarded to a neighbor with the optimal decision rule exploiting the local information. Similarly, the discovery packet is forwarded to the subsequent nodes, until it reaches the destination, hopefully with the smallest number of hops.

In this scheme every condition, which is used for finding the next hop, is first checked for the immediate neighbors, and if no neighbor satisfies it the two-hop neighbors are checked. The next hop is selected based on the similarity of the immediate/two-hop neighbor to the destination. The similarity means that either the node has the same AS Number (ASN) as the destination or it shares the country code with the destination. Otherwise, the more connected immediate/two-hop neighbor determines the next hop. The connectivity is expressed by the node degree. Once the discovery packet reaches a node sharing the country code with the destination, the destination's ASN is sought within the particular country. During the discovery process, an online path optimization mechanism is employed to reduce the path length of the searched path by utilizing 2-hop neighborhood information. The discovery mechanism does not consider an already visited neighbor as a next node to avoid loops in the final path. An example of this mechanism is provided in Fig. 2.

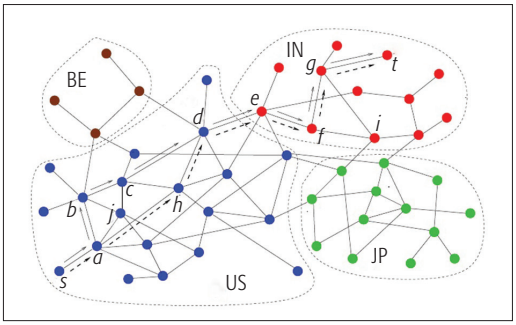

FIGURE 2. Example of route discovery mechanism. This mechanism first finds a path *s-a-b-c-d-e-f-g-t*. The path optimization mechanism attempts to reduce the length of this path at each node. This mechanism produces a shorter path, *s-a-h-d-e-f-g-t*. The two-hop neighbor information of *d* contains *a*. As a consequence, *b* and *c* are replaced by *h*. The loop avoidance mechanism prevents retracing the already visited node *e*, once *f* is reached. This enables the selection of *g* as the next node, which has the same preference as *e*. Since *b* has an option to choose the next neighbor among *c* and *j*, a random selection is applied to pick *c*.

Distributed Compact Routing (DCR): In [3], another distributed unicast compact routing scheme, DCR, based on the centralized scheme AGMaNT (ref. [4] in [3]), was proposed.

In this scheme, each node in the network picks a color from a small set of colors at the same time. All nodes share a hash function that maps identity/address to an element of the color set.

The *vicinity* of a node is the minimal set of close nodes that contains at least one node of each color. The size of this set can be proven to be proportional to the number of colors. Each node stores a direct route to the nodes in its vicinity. Moreover, for all other nodes having a hash value equal to its color, it stores the address of/route to a landmark that has that node in its vicinity. When a node has to forward a packet, it first checks whether it has a route based on its identifier. If not, the hash is determined and the packet is forwarded to a node in the vicinity with the same color. The routing path via a landmark

[13] Among the proposed compact routing schemes, DCR and GCMR are locator-independent.

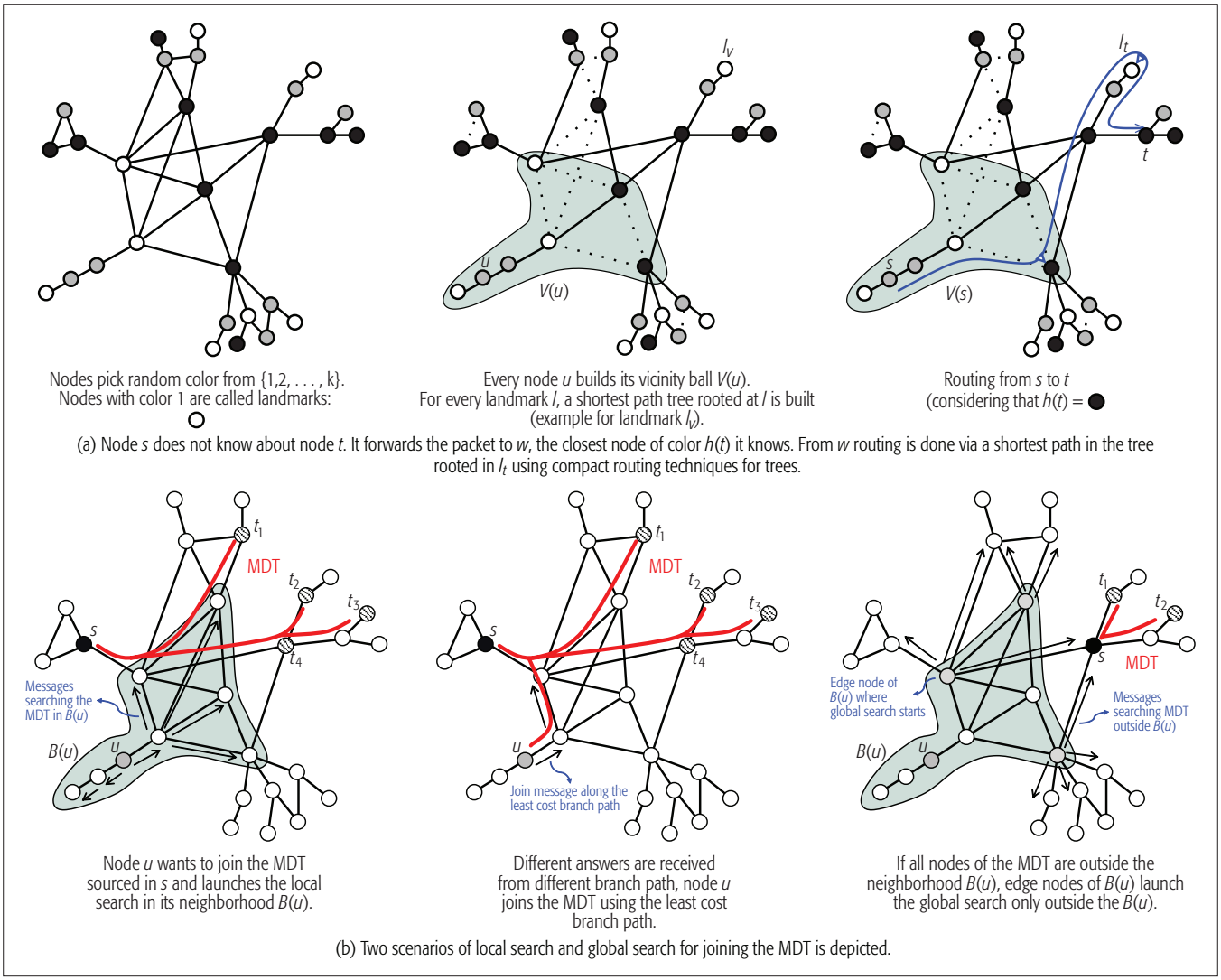

FIGURE 3. Different steps in DCR are depicted in a); b) illustrates an example of GCMR.

has to be encoded in the header to allow routing, which imposes storing a compressed path. To this end, we use a compact routing scheme dedicated to trees. Figure 3a visualizes the steps of DCR.

Greedy Compact Multicast Routing (GCMR): GCMR [4] is a multicast scheme that minimizes the RT size of each node at the expense of paths with relatively small deviation from the optimal stretch, and higher communication cost compared to shortest path tree. GCMR minimizes the storage of routing information by requiring only neighbor-related information. Thus, it does not rely on the construction of global structures such as sparse covers or trees. To limit the communication cost, the routing information needed to reach a given multicast source is acquired by means of an incremental two-stage search process. First, the joining node searches nodes belonging to the multicast tree in its neighborhood (local search); in case of failure, the search is then continued over the remaining unexplored topology (global search). The request message includes a path budget that is used to limit the distance traveled by requests in the local search. Starting from the joining node, this value is decremented in every intermediate node.

The joining node sends a request to its neighbors and starts a timer. The neighbors propagate the message following a split horizon[14] until it reaches a node that is in the MDT or is an edge node of the neighborhood (i.e., path budget reaches 0). The receiving node sends back a reply indicating whether it belongs to MDT or not. Based on this information all the nodes along the path to the joining node compute their path cost. At the joining node, if the timer expires and no reply message is received, it triggers the global search. The joining node sends a request message directly to each edge node. This is possible because during the local search, the received reply messages include the identifier of the edge nodes that initiated them. Additionally each intermediate node keeps an active interface toward each edge node. In the global search the path budget is set to the graph diameter and the waiting timer to a value that prevents waiting indefinitely. Figure 3b illustrates an example of GCMR.

GCMR is adaptive and the adaptability mech-

[14] Split horizon is a method to prevent a routing loop in a network. The principle of this method is to never send back the routing information of a packet in the direction from which it was received.

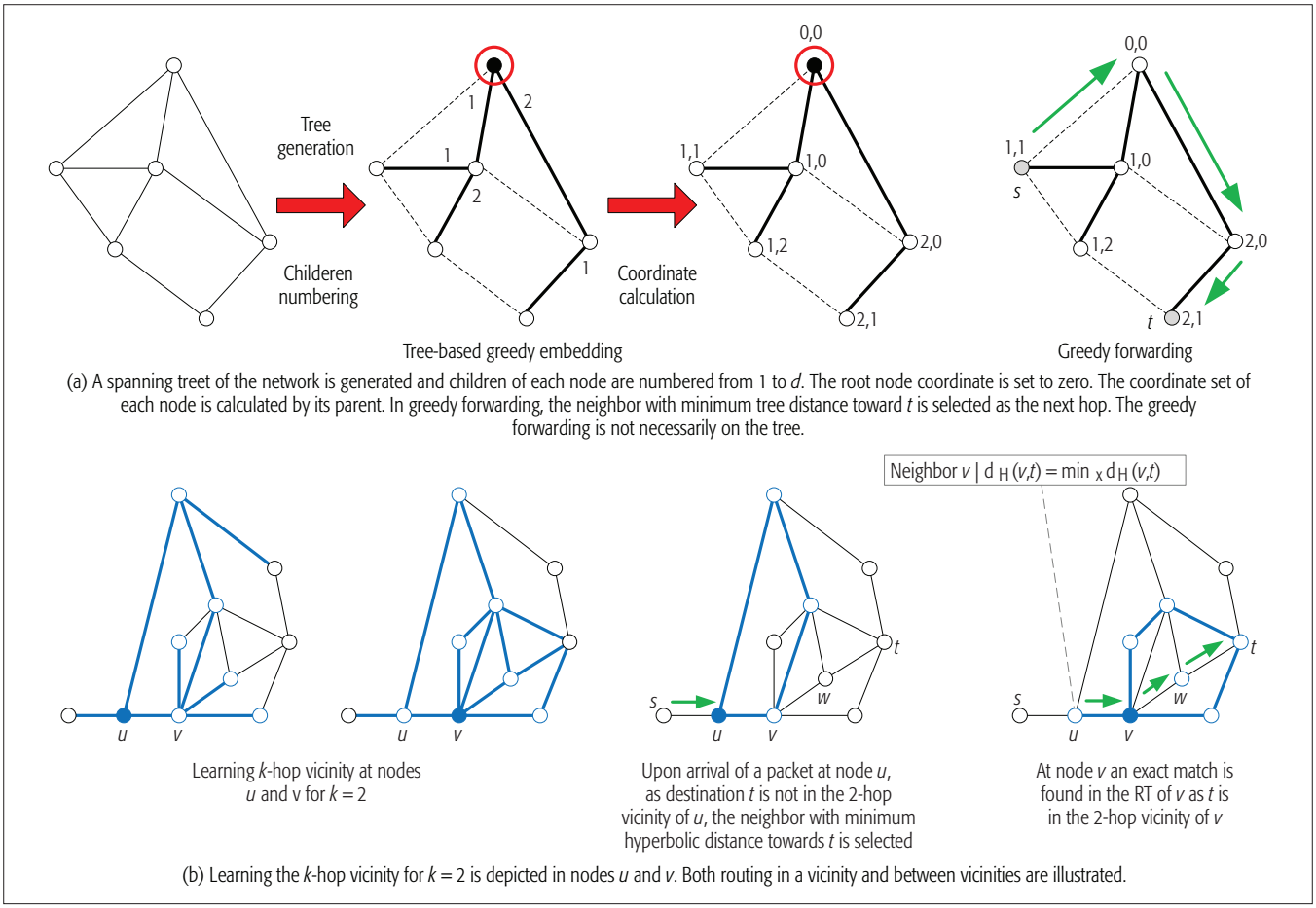

FIGURE 4. Variants of geometric routing: a) an example of GTR embedding and forwarding; b) the principles of GCLS

anism, which is based on a modified two-stage search process, is initiated by the upstream node with respect to the point of change.

Geometric Routing: Geometric routing provides an alternative mechanism trading-off dynamics with increased memory efficiency.[15] We have investigated three classes: geometric tree-based greedy routing, word-metric-based greedy routing, and geometric coordinate-labeling scheme. The first two schemes are based on a tree structure and follow similar procedures. All schemes rely on embeddings into metric spaces to assign coordinates to nodes that are used as locators to perform point-to-point routing decisions in this space.[16]

Geometric Tree-Based Greedy Routing (GTR) and Word-Metric-Based Greedy Routing (WMGR): GTR [5] and WMGR [6] comprise two components: greedy embedding and greedy routing/forwarding. The greedy embedding scheme finds a mapping between nodes and coordinates in a metric space in such a way that there is always a distance-decreasing neighbor toward any destination in the network. These coordinates are then used by the forwarding component to forward the packets toward the intended destinations. Knowing the coordinates of the neighbors, in order to forward an incoming packet, the distance between every neighbor and the packet's destination is calculated. The neighbor with the maximal decrease in the distance is selected as the next hop.[17] This scheme is referred to as greedy routing/forwarding because in each step, the node with maximum decrease in the distance is selected.

In GTR and WMGR, coordinates are determined based on a network spanning tree. While GTR calculates coordinates based on the path from the tree root to each node, WMGR relies on a word-metric space (WMS) that is generated by an algebraic group, where the distance function between two elements is the shortest path length of the corresponding vertices in the Cayley Graph of the group. Considering the free group[18] with a generating set S, the embedding in

[15] This class does not provide locator-independent addresses, by design. The independence is a key feature of the proposed compact routing schemes (DCR, GCMR).

[16] Geometric routing and compact routing themselves are not mutually exclusive. Indeed, the coloring technique used in compact routing might in fact rely on geometric routing rather than typically proposed Cowen landmark routing.

[17] Greedy embeddings guarantee that packets that are forwarded following the distance-decreasing policy will eventually reach their destination. In the absence of such embeddings, packets may reach a local minimum, a node that is closer to (but different from) the destination than any of its neighbors. Alternate solutions such as face routing techniques also enable to bypass the local minima. However, these techniques require that the network topology is planar or planarized, which may not always be feasible.

[18] In mathematics, the free group over a given set S comprises all expressions (words or terms) that can be generated from the members of S.

	Stretch	Memory complexity: input	Memory complexity: output	Communication complexity
DCR	5	$\tilde{O}\sqrt{n}$	$\tilde{O}\sqrt{n}$	$O(n^{3/2})$
GCMR	$O\left(\frac{D(G)+1}{2}\right)$	$O(\Delta(G)\cdot\log(n))$	$O(h\cdot\log(n))$	$2m$
GTR	$O(\log(n))$	$O(\Delta(G))$	$O(\Delta(G)\cdot\log^2(n))$	$O(n+\log(n)\cdot m)$
WMGR	$O(\log(n))$	$O(\Delta(G))$	$O(\Delta(G)\cdot\log2(n))$	O(n+log(n)· m)
GCLS	(2, 2δ)	$\tilde{O}\left(n^{1/2}\right)$	$\tilde{O}\left(n^{1/2}\right)$	$\tilde{O}\left(n^{3/2}\right)$
BGP	1	$O(D(G)\cdot n\cdot(n-1)\cdot\log(n))$	$O(D(G)\cdot n\cdot\log(n))$	$\tilde{O}\left(n^2\cdot(n-1)\cdot\text{polylog}(n)\right)$

TABLE 2. Comparison stretch — memory — communication cost. In this table, m stands for the number of links, n is the number of nodes and $D(G)$ is the diameter of graph G. $\Delta(G)$ represents the maximum nodes degree and h is the size of MDT in multicast routing. δ is the Gromov delta which measures the deviation of the graph from tree—likeness.

Within EULER we performed in-depth evaluation of the schemes explained later on large-scale scale-free networks and compared the results/identified trade-offs with BGP since it is the only routing protocol that has actually been applied in a large-scale scale-free network.

WMGR involves mapping the network spanning tree into the Cayley Graph of the free group. The required steps for calculating the embedding in both schemes are:

- A rooted spanning tree of the network is generated.
- In GTR, the root node sets its coordinates to zero while in WMGR, knowing that S is an alphabet of symbols s_i and a word is a sequence of these symbols, the root is assigned a label that is an empty word of the group (identity).
- In GTR, each node numbers its children from 1 to d and calculates their coordinate sets (CS) by putting the child's assigned number in place of the first zero coordinate in its own CS. Similarly, in WMGR, every node v assigns to its i-th child, a label that is the concatenation of its label and s_i.

For GTR, we propose to use tree-distance as a metric, which is the hop-count on the tree between two nodes and is calculated as follows:

- The closest common ancestor to both nodes is found.
- The hop-count of each node to the ancestor is counted.
- The sum of these hop-counts determines the tree-distance between them.

Figure 4a depicts the greedy embedding and greedy forwarding in GTR.

In WMGR, given the labels of two vertices u and v, we distinguish between a common prefix (the set of first symbols that are equal) and the suffixes (the rest of symbols). The distance between u and v is the length (number of symbols) of the word composed by the concatenation of these two suffixes.

Simulation experiments proved that both schemes perform equally well as other greedy embedding-based schemes in terms of stretch but better in terms of coordinate memory scaling.

In these schemes, adaptivity with respect to changing topology is achieved via an on-demand discovery component to bypass failing elements. The latter can be proactively activated, or can be executed upon failure detection. If these techniques are not applied, re-convergence of the supporting spanning tree is needed, resulting in coordinates re-calculation for a (sub-)tree of the topology.

Geometric Coordinate-Labeling Scheme (GCLS): GCLS [7] is the extension of the previously explained geometric routing schemes. It uses k-hop neighborhood information instead of the default 1-hop neighborhood. GCLS relies on hyperbolic geometry in which coordinate calculation is based on a distributed process where all nodes send information to their neighbors. Coordinates are then derived from round-trip times [8] transformed into hyperbolic distances.

In order to dynamically populate the routing tables with entries pointing to the calculated coordinates, this scheme follows a modified distance-vector algorithm. Each node maintains a vector of distance from itself to all nodes within k hops. Note that the calculated distances in this modified version are based on the hyperbolic distance.

The nodes within maximally k-hop distance form a k-hop vicinity. In the RT of each node in a vicinity, there is an exact match for each destination node that belongs to the same vicinity.

The scheme combines exact match lookup (locally reachable vicinities) and greedy forwarding (remotely accessible vicinities). Upon receiving a packet, first it is checked if an exact match is found in the local RT. In case of a miss, the hyperbolic distance between every neighbor and the destination is calculated and the neighbor with minimum distance is selected as the next hop.

A practical scheme is obtained by:

- Limiting greedy decisions for distant routing to $\sqrt{n}$ interconnected nodes (discovered during coordinate construction) referred to as landmarks (locally selected when exchanging local routing information).
- Let local routing decision based on decision derived from k = 2 neighborhood information.

Figure 4b depicts the principles of GCLS.

Routing Schemes for DTN and P2P Networks

Delay Tolerant Network (DTN): The concept of DTN was introduced initially in the research efforts made for Interplanetary Internet. However, today it is known that a similar concept can be applied to many other networks called "challenged networks." The main characteristics of such networks are frequent disruption, sparse network density, high error rate, delay and mobility. Routing in such networks is quite challenging. The authors in [9] survey many of the recent routing schemes for DTNs. A known scheme in such networks is epidemic routing, in which a message is replicated to all neighbors except the one on which the message arrived. Different improvements to this routing and hybrid schemes such as epidemic routing combined with network coding are described in [9]. Similar to DTNs, wireless mobile ad hoc networks are considered infrastructureless and dynamic in nature. Stochastic routing is considered to be a promising paradigm in such networks. In this routing the next hop in a path is selected according to a probability distribution. Many factors such as load, residual energy and forwarding cost can be used to influence this distribution.

Peer-to-Peer (P2P) Overlay Network: P2P networks, initially introduced as a simple music sharing application, are today responsible for a significant share of Internet traffic. P2P overlays are logical topologies on top of the physical networks which can be built dynamically. These networks are highly scalable, resilient and self configurable, which motivates their widespread use. The authors in [10] survey several algorithms and mechanisms considered in P2P overlay networks. One of the interesting concepts considered in P2P is applying a DHT structure on top of the overlay. Using this structure, (key, value) pairs are stored in a DHT, and all participating nodes can retrieve the value associated with a key efficiently.

Comparative Analysis and Identified Trade-Offs

Within EULER we performed in-depth evaluation of the schemes explained later on large-scale scale-free networks and compared the results/identified trade-offs with BGP since it is the only routing protocol that has actually been applied in a large-scale scale-free network.

Table 2 compares the upper bounds of the performance metrics characterizing routing algorithms. These complexities correspond to the case of scale-free networks. Then, we detail the trends/trade-offs in different routing components, identified through excessive simulation/emulation experiments, which should enable more careful conclusions regarding the applicability of the schemes to large-scale/complex networks.

Route Discovery (RD)

This scheme could discover near-optimal paths in most cases, even when a significant number of links/nodes are suppressed. Incorporating a moderate global knowledge about the network structure — group membership — induces a steep improvement in performances. The identified trade-offs are:

P2P networks, initially introduced as a simple music sharing application, are today responsible for a significant share of Internet traffic. P2P overlays are logical topologies on top of the physical networks which can be built dynamically. These networks are highly scalable, resilient and self configurable, which motivates their widespread use.

- Group information in the packet and at each node decreases the search area.
- The online path optimization mechanism significantly reduces the discovered path length.
- Topological information needed at each node depends on the node degree of its neighbors.

DCR

Simulation results indicate that the actual stretch and the RT size of DCR are far better than the theoretical ones [3]. Comparative evaluations of different algorithms indicate that exploiting topological properties helps improve the performance in some approaches [11]. For instance, CLUSTER using power-law graphs properties, is efficient on all criteria if the network has small-world properties.[19]

However, the performance of this algorithm degrades drastically in other networks (e.g. unit disk graph[20]) [11]. On the contrary, DCR has a trade-off between communication cost/stretch independent of the considered graph. In different topologies, DCR achieves a communication cost almost 10 times smaller than BGP with an average stretch of less than 2 and a maximum number of entries 10 times smaller than BGP.

GCMR

Simulation/emulation results confirm that GCMR, compared to state-of-the-art such as PIM, SPT and ACMR [12], achieves the lowest memory space for storing the routing information, a minimum stretch factor increase (w.r.t. the optimal one), and the lowest recovery/convergence time in case of failure, while further improvements in terms of communication cost are required [4]. Regarding identified trade-offs, additional information in the RTs allows a large reduction in the communication cost, decreases the stretch, and allows a low reduction in the convergence time.

GTR-WMGR

Simulation/emulation outcomes support the memory-advantage of both GTR and WMGR, but clearly indicate the resulting cost in the recovery domain [5, 6]. Similar to other greedy routing schemes, the RT size is bounded by the maximum vertex degree. On scale-free graphs, these schemes achieve good trade-offs among different metrics: they are scalable in storage space, they are succinct (labels are of size $O(\text{polylog}(n))$ bits), and they have a bounded low-stretch. The identified trade-offs in case of failures are:

[19] In a network with small-world properties, the typical hop-count between two randomly chosen nodes grows proportionally to the logarithm of the number of nodes in the network.

[20] A unit disk graph is the intersection graph of circles of unit radius in the Euclidean plane. In this graph, each vertex corresponds to a circle and two vertices are connected by an edge if and only if the corresponding circles intersect.

The experiments performed so far indicated that the proposed schemes have interesting characteristics in terms of memory usage, stretch and convergence behavior, which make them promising schemes for large-scale complex networks. Indeed, there are some open problems that require further research.

- A potentially high number of affected paths, with a generally low convergence time.
- Protection: fast recovery with no communication overhead, but high stretch.
- Restoration: high convergence-time/communication-cost with potentially low stretch.

Schemes such as GCLS, GCMR and RD show adaptability to failures by only requiring re-computation of the routes affected by the failure. The number of affected routes is proportional to the centrality of the failing entity. GTR and WMGR provide protection capability to overcome pre-determined failure patterns, and if no protection exists they re-calculate the coordinates of the affected nodes. DCR, on the other hand, does not provide dedicated processing for information state changes and requires the full re-computation of the routing tables.

Exploiting the topological properties of scale-free networks can improve the performance of several compact routing schemes [11]. This was also confirmed in GTR and WMGR schemes as the tree construction method (i.e., selection of the maximum degree node as the root and construction of a breadth-first-search tree) does not generate deep branches due to the short average distance between nodes in scale-free networks. This minimizes the memory requirements for coordinate representation and enables shorter paths [13].

In all schemes, a scalable mapping system to bind node identifiers to node locators is required. An option is to use DNS-like servers for these name-to-address/address-to-address resolutions. The main identified trade-off is between communication-cost and convergence-time. The slower the polling scheme relative to the mapping service, the smaller the communication-cost but the longer the convergence time.

The proposed schemes have different packet forwarding processes. While GTR, WMGR and RD replace the lookup with more computation in the forwarding plane, GCLS and DCR look up the next hop of a packet from the RTs.

Conclusion and Future Directions

We presented an overview of potential design dimensions of routing protocols, the routing functionality and existing routing schemes. The focus of the article was mainly on advances in compact routing and geometric routing. Each of the studied routing schemes has its own set of functional and performance related characteristics, as described in Tables 1 and 2. These result in careful trade-offs to be made, without one-size-fits-all. Compact routing and geometric routing have roughly similar performance characteristics in which geometric coordinates encode similar locational information as the compact forwarding tables in compact routing. Lookup logic has been largely optimized in current routers, while greedy forwarding might require changes in typical forwarding logic and hardware.

Through analyzing experimental/analytical results performed so far, we identified the main open challenges:

•One cause of absence of an alternative to BGP is that the design of many routing systems (mainly path-vector scheme enhancements) tends to follow the same approach as the one pursued by BGP. To overcome this, clean-slate approaches should be investigated.

•Most investigated schemes increase performance by decreasing functionality. For example, all previously listed schemes improve scalability in terms of memory. However, the same level of policy as in BGP is not supported in any of them.

•The main difference in the scheme discovery process results from the exchange of routing information: pull vs. push. All alternatives use a distance metric/spatial routing metric that subdivides between local and global metrics and between metrics derived from the topology properties (e.g. node degree) and universal metrics. These dimensions are tightly related, and our results corroborate that schemes such as BGP, which is independent of global or link metrics and is driven by local policy decisions, will be challenging to replace as long as the Internet domains are operated organically.

•From the experimental perspective, due to the increasing level of path-processing granularity combined with a larger parameter space, deriving common path characteristics to obtain a representative policy model together with the AS relationships remains challenging.

The experiments performed so far indicated that the proposed schemes have interesting characteristics in terms of memory usage, stretch and convergence behavior, which make them promising schemes for large-scale complex networks. Indeed, there are some open problems that require further research:

•Many of the schemes rely on a tree construction. It is thus appropriate to further investigate multi-path routing via independent trees in order to extend these schemes with fault-tolerance and load balancing [14]. Additionally, using multiple trees is a starting point for enabling routing policy in these schemes.

•Many of the schemes require a mapping system. Despite the research efforts (mainly in LISP [15]) a scalable, secure and highly reliable mapping system with fast convergence is still missing.

•The proposed schemes find applicability in upper layers (IT/computing systems, information/file systems) when the number of entities reaches at least 10^{10}. Concretely, content-centric networking (CCN) is one paradigm that can benefit from the proposed geometric routing schemes. Using these schemes, an efficient and scalable content-based forwarding is possible, which was demonstrated in [7]. If capacity saving remains a key objective, integrating multicast benefits with CCN should be further investigated.

•Although the previous routing schemes are proposed for networks far from complex networks, it is an interesting research direction to investigate the applicability of such routing schemes in large-scale scale-free networks. Particularly, schemes such as stochastic routing may

be a promising alternative if parameters such as network load are used in the calculation of probability distribution. In this way an adaptive load balancing mechanism can be achieved. P2P networks as a potential data distribution paradigm of the future Internet require further investigation to improve aspects such as security, dynamicity, redundancy and load balancing [10].

•Finally, it is challenging to translate schemes/algorithms into protocols, and it is a research topic on its own.

Schemes such as stochastic routing may be a promising alternative if parameters such as network load are used in the calculation of probability distribution. In this way an adaptive load balancing mechanism can be achieved. P2P networks as a potential data distribution paradigm of the future Internet require further investigation.

Acknowledgment

This work is partly funded by the European Commission through the EULER project (Grant 258307), part of the Future Internet Research and Experimentation (FIRE) objective of the Seventh Framework Programme (FP7), the UGent BOF/GOA project 'Autonomic Networked Multimedia Systems' and the Spanish Government (GIROS, TEC 2015-66412-R).

References

[1] F. Coras *et al.*, "Locator/ID Separation Protocol (LISP) Impact," 2016.
[2] P. Godfrey *et al.*, "Stabilizing Route Selection in BGP," *IEEE/ACM Trans. Networking (TON)*, vol. 23, no. 1, 2015, pp. 282–99.
[3] C. Gavoille *et al.*, "On the Communication Complexity of Distributed Name-Independent Routing Schemes," *Int'l. Symposium on Distributed Computing*, Springer, 2013, pp. 418–32.
[4] D. Careglio *et al.*, "Development and Experimentation towards a Multicast-Enabled Internet," *Proc. Computer Communications Workshops (INFOCOM WKSHPS)*, 2014, pp. 79–84.
[5] S. Sahhaf *et al.*, "Experimental Validation of Resilient Tree-Based Greedy Geometric Routing," *Computer Networks*, vol. 82, 2015, pp. 156–71.
[6] M. Camelo *et al.*, "Geometric Routing with Word-Metric Spaces," *IEEE Commun. Lett.*, vol. 18, no. 12, 2014, pp. 2125–28.
[7] S. Sahhaf *et al.*, "Experimentation of Geometric Information Routing on Content Locators," *Proc. 2014 IEEE 22nd Int'l. Conf., Network Protocols (ICNP)*, 2014, pp. 518–24.
[8] T. E. Ng and H. Zhang, "Predicting Internet Network Distance with Coordinates-Based Approaches," *Proc. Twenty-First Annual Joint Conf. IEEE Computer and Communications Societies*, INFOCOM 2002, vol. 1, 2002, pp. 170–79.
[9] Y. Cao and Z. Sun, "Routing in Delay/Disruption Tolerant Networks: A Taxonomy, Survey and Challenges," *IEEE Commun. Surveys & Tutorials*, vol. 15, no. 2, 2013, pp. 654–77.
[10] A. Malatras, "State-of-the-Art Survey on P2P Overlay Networks in Pervasive Computing Environments," *J. Network Computer Applications*, vol. 55, 2015, pp. 1–23.
[11] C. Gavoille *et al.*, "Brief Announcement: Routing the Internet with Very Few Entries," *Proc. 2015 ACM Symposium on Principles of Distributed Computing*, 2015, pp. 33–35.
[12] I. Abraham, D. Malkhi, and D. Ratajczak, "Compact Multicast Routing," *Int'l. Symposium Distributed Computing*, Springer, 2009, pp. 364–78.
[13] S. Sahhaf *et al.*, "Efficient Geometric Routing in Large-Scale Complex Networks with Low Cost Node Design," *IEICE Trans. Commun.*, vol. 99, no. 3, 2016, pp. 666–74.
[14] R. Houthooft *et al.*, "Robust Geometric Forest Routing with Tunable Load Balancing," *Proc. 2015 IEEE Conf. Computer Communications (INFOCOM)*, 2015, pp. 1382–90.
[15] A. Anisul and H. Flinck, "A Compact Routing Based Mapping System for the Locator/ID Separation Protocol (LISP)," *Int'l. J. Computer Applications*, vol. 127, no. 5, 2015, pp. 1–8.

Biographies

Sahel Sahhaf received a M.Sc. degree in information technology from Sharif University of Technology (SUT), Tehran, Iran in 2010. In November 2010 she joined the Internet Based Communication Networks and Services (IBCN) research group at Ghent University as a Ph.D. student. She has focused on novel network architecture for next generation Internet and also concentrated on network function virtualization and service function chaining.

Wouter Tavernier received a M.Sc. in computer science in 2002 from Ghent University (Belgium). After a two-year period as a business analyst at Accenture Belgium, he joined the Internet-Based Communication Networks and Services group of Ghent University in 2005 to research Future Internet topics, including resiliency of (Carrier) Ethernet and IP networks, geometric routing and the application of (machine) learning techniques in the context of routing and switching. This research led to a Ph.D. degree in computer science engineering from Ghent University (Belgium) in 2012. His current research focuses on software-defined networks and service function chaining.

Dimitri Papadimitriou started at Alcatel in 2000, working on multi-layer traffic-engineering research for the Corporate Research Center. In 2003, he joined the Research and Innovation Department dedicated to network distributed control and routing algorithms. Since 2007, he has been working as a senior researcher at Nokia Bell Labs, currently in the Network Algorithmic Analytics Control and Security Research Lab. His research interests include network optimization and algorithms, optimization under uncertainty, and computational intelligence. He has led several EU FP7 research projects over last 10 years.

Davide Careglio received the M.Sc. and Ph.D. degrees in telecommunications engineering from the Technical University of Catalonia (UPC), Barcelona, Spain, and the Laurea degree in electrical engineering from Politecnico di Torino, Torino, Italy. He is an associate professor with the Department of Computer Architecture, UPC. His research interests include algorithm and protocol design, modeling, and optimization in communication networks.

Christian Glacet received his M.S. degree in computer science from the University of Bordeaux (France) in 2010, and his Ph.D. degree in computer science from LaBRI Bordeaux in 2013. He is currently a lecturer at the University of Bordeaux and a member of the LaBRI. His research interests include distributed computing, graph theory and routing.

David Coudert has been a senior research scientist at Inria Sophia Antipolis since 2002, and the scientific leader of COATI, a joint project-team between Inria and the I3S (CNRS, UNS) laboratory. He received his Ph.D. degree in computer science from the University of Nice Sophia Antipolis in 2001 and his habilitation in 2010. His research interests include algorithmic graph theory, combinatorial optimization and operations research with communication networks as a main application area (routing, fault tolerance, reliable design, and so on). He is also working on optimization problems from structural biology and transportation networks. He is on the editorial board of *Networks* (Wiley) and *Discrete Applied Mathematics* (Elsevier).

Nicolas Nisse received his M.S. and Ph.D. degrees in computer science from the University of Paris-Sud (France) in 2004 and 2007, respectively. He was a postdoctoral research associate in the Departamento de Ingenieria Matematica (University of Chile, Santiago, Chile) in 2007-08, and at Inria Sophia Antipolis (France) in 2008-09. Since 2009, he has been a full time researcher at Inria Sophia Antipolis. He received his habilitation à Diriger la Recherche (HdR) in computer science from theUniversity of Nice Sophia Antipolis in 2014. His research interests include graph algorithms, graph theory, and combinatorial optimization, mainly focusing on the spreading of information/routing in networks and pursuit-evasion games in graphs.

Lluís Fàbrega has been an associate professor in computer science at the University of Girona (UdG) since 2008. He received a degree in telecommunications engineering (1995) and a master's degree in mobile communications (1996) at the Polytechnical University of Catalonia, and a Ph.D. degree in computer science at the UdG (2008). His research interests include the design and performance evaluation of routing, traffic engineering and quality of service mechanisms in the Internet and in connection oriented network technologies. He has worked on several Spanish and EU (Celtic, FP7) research projects, and he has co-authored several papers in journals and international conferences.

Miguel Camelo is a researcher in the Department of Mathematics and Computer Sciences at the University of Antwerp, Belgium. He is an electronic engineer from the University of Ibague, Colombia, 2006, and a master in systems and computer engineering from the University of Los Andes, Colombia, 2010. He received his Ph.D. in computer engineering at the University of Girona, Spain, 2014. His research interests are in the fields of control and management of communication networks and routing in complex networks using group theory, evolutionary algorithms and machine learning. He has worked on several Spanish, Belgian and European research projects.

Pere Vilà is an associate professor in computer science at the University of Girona. He is a computer science engineer (1997) from the Polytechnic University of Catalonia, and he received a Ph.D. in computer science (2004) from the University of Girona. His research interests include new routing algorithms for Future Internet, network protection and robustness, complex networks, network management and control. He has authored or co-authored approximately 50 papers in his areas of interest, served as a member of program committees in several international conferences, and as an associate editor of the *International Journal of Communication Systems*.

Pieter Audenaert received the M.Sc. in pure mathematics and the Ph.D. degree with a focus on theoretical aspects of computer science from Ghent University, in 2000 and 2004, respectively. Currently, he is affiliated with Ghent University/IMEC and works in the field of networks in its broadest sense: from communication networks and logistic networks, to protein-interactions and social networks. To this aim, he specializes in graphs and algorithms with a focus on applying theoretical results in the field of computer science. This entails data-modeling, computational analysis and statistical forecasting.

Didier Colle is a full professor at Ghent University. He received a Ph.D. degree in 2002 and a M.Sc. degree in electrotechnical engineering in 1997 from the same university. He is group leader in the IMEC Software and Applications business unit. He is co-responsible for the research cluster on network modelling, design and evaluation (NetMoDeL). This research cluster deals with fixed Internet architectures and optical networks, Green-ICT, design of network algorithms, and techno-economic studies.

Piet Demeester is a professor on the faculty of engineering at Ghent University. He is head of the research group 'Internet Based Communication Networks and Services' (IBCN, Ghent University) and leads the Internet Technologies Department of the strategic research center iMinds. In 2008 he was named a Fellow of the IEEE "for contributions to optical communication networks and technologies."