

Dear Author,

Please, note that changes made to the HTML content will be added to the article before publication, but are not reflected in this PDF.

Note also that this file should not be used for submitting corrections.



The Journal of Systems and Software

journal homepage: www.elsevier.com/locate/jss

Alleviating the topology mismatch problem in distributed overlay networks: A survey

Q1 Vassilis Moustakas^a, Hüseyin Akcan^b, Mema Roussopoulos^c, Alex Delis^{c,*}^a Amazon Development Center Germany GmbH, Marcel-Breuer Str. 12, 80807 Munich, Germany^b Department of Software Engineering, Izmir University of Economics, 35330 Balcova, Izmir, Turkey^c Department of Informatics and Telecommunications, University of Athens, Athens 15703, Greece

ARTICLE INFO

Article history:

Received 23 July 2015

Revised 30 September 2015

Accepted 17 November 2015

Available online xxx

Keywords:

Distributed systems

Peer-to-peer

Overlay network

Topology mismatch

Topology awareness

ABSTRACT

Peer-to-peer (P2P) systems have enjoyed immense attention and have been widely deployed on the Internet for well over a decade. They are often implemented via an overlay network abstraction atop the Internet's best-effort IP infrastructure. P2P systems support a plethora of desirable features to distributed applications including anonymity, high availability, robustness, load balancing, quality of service and scalability to name just a few. Unfortunately, inherent weaknesses of early deployments of P2P systems, prevented applications from leveraging the full potential of the paradigm. One major weakness, identified early on, is the *topology mismatch problem* between the overlay network and the underlying IP topology. This mismatch can impose an extraordinary amount of unnecessary stress on network resources and can adversely affect both the scalability and efficiency of the operating applications. In this paper, we survey over a decade's worth of research efforts aimed at alleviating the topology mismatch problem in both *structured* and *unstructured* P2P systems. We provide a fine-grained categorization of the suggested solutions by discussing their novelty, advantages and weaknesses. Finally, we offer an analysis as well as pictorial comparisons of the reviewed approaches since we aim to offer a comprehensive reference for developers, system architects and researchers in the field.

© 2015 Published by Elsevier Inc.

Q2 1. Introduction

Peer-to-peer (P2P) networks are self-organizing, distributed systems where participating nodes, called *peers*, act as both resource providers and resource consumers in contrast to the conventional client-server model where nodes undertake specific roles. For over a decade, P2P networks have been widely deployed and have enjoyed immense popularity from Internet communities, primarily because of the great number of features they offer to distributed applications built atop them. Such diverse features include: high availability and robustness, load-balancing, quality of service, scalability, decentralized administration, and anonymity.

The peer-to-peer paradigm gave impetus to two "killer" applications: file-sharing and Internet telephony. The Napster file-sharing system was widely acknowledged as the "fastest growing Internet application ever" in 2001 when it topped 26 million users sharing over 80 million songs. Other file-sharing applications followed

suit, including Gnutella and Limewire enjoying 3 million concurrently-connected peers, as well as BitTorrent connecting over 150 million monthly users by January 2012 (BitTorrent, 2012). P2P telephony saw explosive growth with the advent of Skype. Since its introduction in 2003, Skype has become extremely popular with more than 650 million users in 2011 (TELECOMPAPER BV., 2011) and an astonishing 50+ million concurrently-online users (Mercier, 2013). Moreover, it has consistently eroded the traffic handled by traditional telephony carriers by slicing away a staggering 214 billion minutes of pertinent traffic in 2013 alone (TeleGeography/PriMetrica, 2014).

Apart from the above, a startling number of diverse and successful applications have been built based on P2P architectures. Some of them include: distributed search engines (YaCY Project, 2012), distributed data-storage systems (Kubiatowicz et al., 2000; Bolosky et al., 2000; Dabek et al., 2001; Druschel and Rowstron, 2001; Adya et al., 2002; Muthitacharoen et al., 2002; Grönvall et al., 2010; Alima et al., 2003), Web caches, archives and publishing systems (Iyer et al., 2002; Bayardo et al., 2002; Waldman et al., 2000; Waldman and Mazières, 2001), messaging and dissemination applications (Microsoft, 2012; Papadimitriou and Delis, 2008), event-notification infrastructures (Rowstron et al., 2001; Castro et al., 2002c; Alima et al., 2003), naming services (Cox et al., 2002), censor-resistant

* Corresponding author. Tel.: +302107275212; fax: +302107275213.

E-mail addresses: vsmoustakas@gmail.com (V. Moustakas), huseyin.akcan@ieu.edu.tr (H. Akcan), mema@di.uoa.gr (M. Roussopoulos), ad@di.uoa.gr (A. Delis).

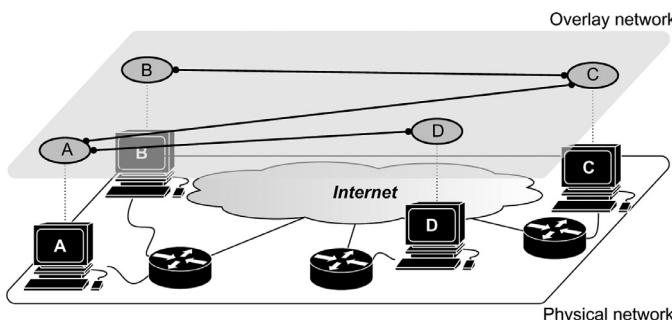


Fig. 1. An example overlay network.

39 stores (Clarke et al., 2000) and lately, even cloud-based platforms
40 (Mondéjar et al., 2013).

41 P2P systems are implemented using *overlay networks*. An over-
42 lay network, is a virtual system of nodes featuring logical intercon-
43 nects (or links) created above an existing network; overlays provide
44 an abstraction that enables the implementation of efficient, fully dis-
45 tributed, application-layer services such as routing messages to des-
46 tinations that are not known in advance or offering QoS guarantees
47 (i.e., in content-distribution) over best-effort infrastructures. Overlay
48 nodes communicate through *virtual connections* each of which may
49 correspond to a path of possibly many physical links in the underly-
50 ing network. Fig. 1 illustrates a simple four-node overlay constructed
51 over a wide-area network.

52 The single key issue that determines the efficiency of an overlay
53 network, is how well the overlay maps to the underlying network
54 topology on which it “rests”. Consider two nodes¹ that are connected
55 with each other via a path of overlay links. If the application run-
56 ning on the nodes, generates heavy traffic along the overlay path, it
57 would be beneficial to construct the overlay topology in a way that
58 the number of underlying IP links between these two nodes is min-
59 imized. Should the overlay network be constructed so that it does
60 not match the underlying topology well, the inherent *topology mis-
61 match* creates two major problems. First, the performance of the
62 application per se, can be adversely affected since traffic must flow over
63 a larger, redundant, number of physical hops resulting in poor user
64 experience entailing noticeable latencies or jitter. Second, other ap-
65 plications running on the underlying network infrastructure may be
66 adversely affected as well. Studies have shown that highly popular
67 P2P applications contribute the largest portion of the overall Internet
68 traffic (Saroiu et al., 2002a; Sen and Wang, 2004; Karagiannis et al.,
69 2005), with some reporting that more than 60% of this traffic to be
70 P2P-related (Ferguson, 2006; Schulze and Mochalski, 2009); it was
71 also projected that this traffic would reach 7 Exabytes-per-month by
72 2018 (Cisco, 2014)! This constitutes a major burden for Internet Ser-
73 vice Providers (ISPs) who must route all of this traffic to destinations
74 at the edge of the Internet. If the P2P overlay topology is poorly de-
75 signed, the demand on the Internet’s backbone infrastructure may
76 substantially increase as traffic might have to flow “back and forth”
77 several times between two neighboring ISPs while trying to travel
78 from the source to its destination node in the overlay. Hence, it is crit-
79 ical that P2P networks be laid out in ways that their topology matches
80 the underlying IP topology as closely as possible.

81 For over a decade, researchers have extensively investigated var-
82 ious aspects of the topology mismatch problem. The main objective
83 of this paper is to offer a comprehensive survey of the work carried
84 out in the area and provide a taxonomy of the proposed solutions. We
85 point out synergies, as well as similarities and differences in the pub-
86 lished approaches. Ultimately, our goal is to help readers sift through
87 the voluminous literature, to help them understand the advantages

and disadvantages of each work, and to provide them with enough
88 perspective so that when the need arises, they are able to select,
89 amongst the different approaches, the one that is most suitable for
90 their particular application. The rest of the paper is organized as fol-
91 lows: In Section 2, we provide background on overlay architectures
92 including centralized, decentralized-unstructured and decentral-
93 ized-structured P2P systems. We also formally define the problem of topol-
94 ogy mismatch and offer the rationale behind our evaluation of the
95 techniques discussed in this paper. Work in the area was inspired
96 mostly by the volatile and fully distributed nature of the decentral-
97 ized overlays so we skip centralized systems and in Sections 3 and 4
98 we outline the surveyed research efforts for unstructured and struc-
99 tured P2P overlays, respectively. We conclude in Section 5.

2. Background

In this section, we provide some basic background information on
102 the types of P2P architectures that have been implemented and de-
103 ployed. We also describe and motivate more formally the topology
104 mismatch problem whose various aspects have been tackled over the
105 past several years.

2.1. Peer-to-Peer (P2P) overlay architectures

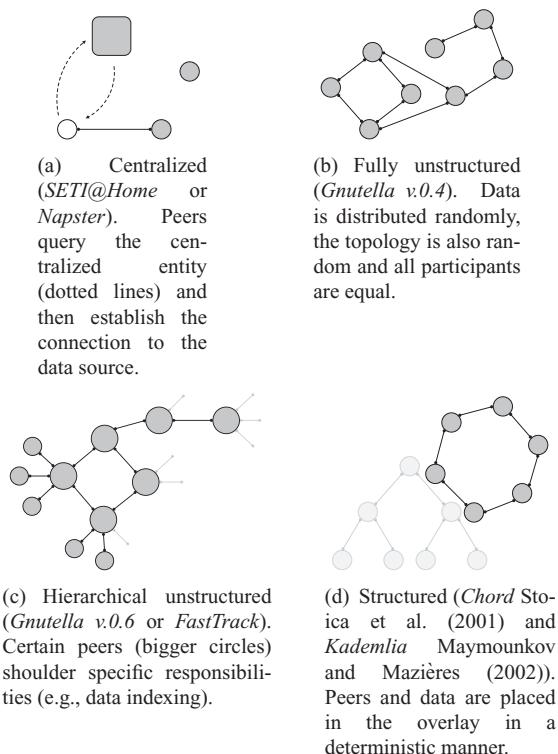
Circa 2000, the first file-sharing P2P-system introduced the ser-
108 vent concept (Gnutella, 2009), a portmanteau that blends the notions
109 of server and client to denote the dual role of a node participating in
110 the P2P-network. Such server-less systems have proven to be able to
111 achieve outstanding aggregate resource capacities as more and more
112 participants join the system without requiring additional expendi-
113 ture for infrastructure. While certain undesirable features such as
114 free-riding (Saroiu et al., 2002b; Adar and Huberman, 2000; Hughes
115 et al., 2005), distribution of illegal content and other socio-technical
116 issues (Hughes et al., 2008) have arisen, P2P systems have contin-
117 ued to gain popularity throughout the years primarily due to their
118 successful file-sharing applications amongst vast numbers of partici-
119 pating users. For over a decade, excitement over the immense poten-
120 tial of P2P systems has generated a flurry of research and develop-
121 ment, resulting in a wide range of popular protocols, networks, and
122 applications. As P2P systems evolved, three main architectures have
123 emerged, namely: (i) centralized, (ii) decentralized unstructured, and
124 (iii) decentralized structured.

Designers of centralized architectures (Fig. 2(a)) were the first to
126 observe that requests for resources (e.g., CPU cycles or popular con-
127 tent) need not be sent to a dedicated server. Instead, such requests
128 could be handled by many hosts that already possessed the resources
129 in question. The downside of this approach was that it required a cen-
130 tralized search directory that inevitably became a single point of fail-
131 ure, a scalability bottleneck or a target of malicious behavior such as
132 DoS attacks.

Napster was the most successful incarnation of the centralized
134 approach in file-sharing (Napster, 2007). It maintained a *central index*
135 server based on file-lists that were peer-provided. Users would
136 query the index server that would provide pointers to the actual
137 content. Thus, by centralizing search while distributing downloads,
138 Napster achieved a highly efficient design that, at its height, attrac-
139 ted 26 million users sharing approximately 80 million songs
140 (JupiterMediaMetrix, 2001). The central indexing component played
141 a vital role in the demise of Napster, as the Recording Industry Associa-
142 tion of America successfully argued in court that this software entity
143 was enabling copyright infringement of the artists’ songs exchanged
144 through the system.

Architectures that distributed both search and resource provi-
145 sion capabilities (Fig. 2(b)) soon followed. In these *decentralized* ap-
146 proaches, resource placement within the overlay topology is random
147 (Yang and Garcia-Molina, 2002); for this reason, such architectures

¹ In P2P networks, the participating nodes are typically user-PCs operating at the edge of the Internet.

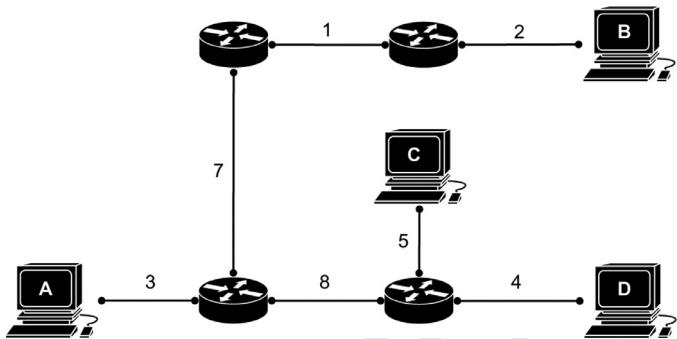
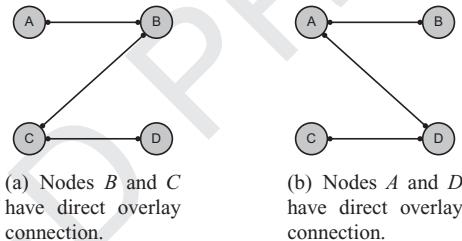
**Fig. 2.** The various P2P architectures.

are referred to as *unstructured*. Key properties of such P2P systems are that they support the inherent heterogeneity of peers, are highly resilient to peer failures, and incur low maintenance overhead while handling the dynamics of peer participation (Stutzbach and Rejaie, 2006). Unstructured systems are also known as *broadcast-based* systems for they use message flooding among peers to propagate search queries. In its first realization, *Gnutella(v.0.4)* became a well-known example of the fully decentralized unstructured approach.

A key disadvantage of the unstructured approach is that message flooding burdens the network since queries travel within the network randomly, visiting nodes that do not have the sought resource and thus wasting bandwidth. To save on bandwidth, unstructured networks typically limit how long a query travels within the overlay via the *time-to-live* parameter (*TTL*). When a query is received by a node, it decreases its *TTL* by one unit. If the *TTL* has not reached zero, the node forwards the query to its neighbors. Otherwise, it drops the query. While this approach prevents queries from visiting the entire network, it does not guarantee that the resource of interest, (e.g., a rare recording) will, ultimately, be found.

To shorten search times, improve scalability, and lighten network load, unstructured configurations evolved into *hierarchical* systems that dynamically assigned indexing functions to special peers, called *ultra- or super-peers* (Fig. 2(c)). Peers with reliable network connections and/or high compute power became *ultra/super-peers* facilitating less-powerful *leaf-nodes* to find resources of interest. Examples of such hierarchical unstructured systems include (Gnutella, 2009), (Kazaa, 2009) and (Skype, 2012). Although the hierarchical approach relieves network pressure as search queries were flooded over the much smaller subset of ultra/super-peers, it fails to address the reduced search scope and inability to locate rare resources.

Decentralized structured schemes were introduced as an answer to the above problems of decentralized unstructured P2P-approaches (Fig. 2(d)). Here, the main objective is to offer a self-organizing infrastructure for large-scale P2P applications (Ratnasamy et al., 2001; Stoica et al., 2001; Rowstron and Druschel, 2001; Zhao et al., 2001; Maymounkov and Mazières, 2002; Rhea et al., 2004). To achieve this,

**Fig. 3.** Interconnection of nodes in the physical level.**Fig. 4.** Two different overlay connection configuration.

they implement a *Distributed Hash Table (DHT)* that maps objects to nodes through a deterministic mechanism. *DHTs* provide a guaranteed bound on the number of overlay routing hops that have to be “traveled” to locate a resource, even in the case when only a *single* copy exists within the system. This requires $O(\log(n))$ hops, compared to the unstructured networks that require $O(n)$ to reliably locate any object. Unfortunately this efficiency does not come without cost. Structured overlays can only support exact-match queries (i.e., queries that identify resources by name) as opposed to content-based retrieval provided by unstructured overlays. This means that unstructured overlays can support more versatile resource location queries such as keyword and partial matching searches. Several popular file-sharing P2P applications have been based on *DHTs* including *Overnet/eDonkey* (Overnet, 2009), *Kademlia/Kad Network* (Maymounkov and Mazières, 2002), and some flavors of *BitTorrent* (Cohen, 2003). Moreover, *DHTs* sparked a flurry of research for many diverse applications, ranging from distributed search engines to event-notification and cloud-based platforms (Kubiatowicz et al., 2000; Rowstron et al., 2001; Mondéjar et al., 2013).

2.2. The topology mismatch problem

One of the major issues that defines the efficiency of an overlay network is its mapping to the underlying physical infrastructure. Recall, in Fig. 1 nodes *A* and *B* are on the same local network, while *C* and *D* are in different networks. The top layer represents the overlay interconnection formed by these four nodes at a higher level. Link connections there, can change as needed by the running environment, without any particular constraint by the underlying physical topology. For instance, assume nodes *A*, *B*, *C* and *D* are connected via the physical network shown in Fig. 3 in which costs are depicted in milliseconds.

When peer *A* sends a message to peer *D*, depending on the overlay configuration, nodes will experience different performance. For example, in the overlay depicted in Fig. 4(a), the message will traverse the following sequence of links in the physical layer (marked with their costs): 3 → 7 → 1 → 2 → 2 → 1 → 7 → 8 → 5 → 5 → 4. In the alternative overlay of Fig. 4(b), the path will be: 3 → 8 → 4. The respective costs are 45 ms and 15 ms. Evidently, the second overlay is

more congruent with the underlying physical network in comparison to the first and, thus, is more efficient.

Ideally, an overlay should be constructed to achieve an *optimal* mapping with its underlying network links, to avoid inefficient states where redundant physical resources are utilized for the operation of the same virtual link in the overlay. The ratio of the actual IP network distance a message travels to reach an object of interest (via overlay routing) to the minimal IP network distance to that object (i.e. through IP) is known as *stretch* or *Relative Delay Penalty (RDP)* (Chu et al., 2000).

The problem of constructing an optimal overlay is referred to as the *topology mismatch problem*, and is formally defined as follows:

Definition 1. Let $V = \{v_1, \dots, v_n\}$ be a set of points denoting the network nodes, $\{v_i, v_j\} \in E$ be the set of unicast distances between nodes v_i and v_j , $G = (V, E)$ be a complete distance graph over V . The topology mismatch problem is to construct a minimal spanning tree, where node degree is restricted to a constant ($k \geq 2$) by the bandwidth of each node v_i .

Early overlay protocol implementations gave little attention to the topology mismatch problem. *Gnutella*, for example, is considered far from scalable (Ritter, 2001) for every peer chooses its neighbors without any knowledge of the underlying network, resulting in mismatches. Queries are flooded over multiple paths in an effort to reach nodes that have the sought content. In addition, pairs of nodes may simultaneously forward the same query to each other in a superfluous mode of operation in this opportunistic and without coordination environment. Ripeanu et al. (2002) showed that, even when 95% of any two nodes are less than 7 hops away from each other, a *flooding-based Gnutella-like* algorithm can generate 330TB/month traffic in a 50,000 node network. A topology mismatch can thus have a serious effect on Internet's routing performance.

Similar problems are observed in decentralized structured schemes as well. Typically, node IDs are assigned *randomly*, resulting in excellent load balancing, scalability, and robustness for the overlay. Unfortunately, this randomness has a negative impact on the *routing locality* of the network. This means that even though the target node can be reached with a logarithmic number of overlay hops, the distance traveled in the underlying network, during the overlay routing process, can be far from optimal.

2.3. Motivation and goal for this survey

A topology-unaware overlay network is able to control the sequence of peers a message traverses before reaching its destination, but it completely ignores how the actual packets are switched at the underlying infrastructure along the overlay path. In particular, a single logical point-to-point link on the overlay typically corresponds to multiple physical links in the underlying layer. Moreover, a link in the physical network often lies on several overlay paths inadvertently increasing the traffic on that link; this is known as the link's stress (Chu et al., 2002). There is also the stochastic phenomenon during which peers depart and others arrive in the network, known as *churn* (Stutzbach and Rejaie, 2006). Even when an approach offers the best possible location when a node joins, churn continuously undermines the effort to maintain the best possible position of the peer in the network.

The topology mismatch problem is exacerbated by the extremely large amounts of traffic generated by highly popular P2P applications (Buford et al., 2009). Addressing this problem remains a fundamental challenge in contributing to the overall efficiency of the Internet. Unfortunately, topology mismatch is known to be an *NP-Hard* problem (Chawathe, 2000; Garey and Johnson, 1979). There is also additional complexity due to the fact that end-to-end latencies demonstrate *triangle inequality violations (TIVs)* which are a consequence of the Internet's structure and routing policies; as such, they will remain a prop-

erty of the Internet for the foreseeable future (Zheng et al., 2005). TIVs affect both network coordinate (Cox et al., 2004; Wong et al., 2005) and positioning (Ng and Zhang, 2001) systems and make the construction of *IP-topology-aware* overlays difficult.

Over the past several years, researchers have investigated an extensive array of heuristic solutions to the topology mismatch problem. The purpose of this survey is to gather and organize this extensive body of produced knowledge. We aspire to assist the reader in better understanding the different approaches suggested and to accomplish this, we present methods for unstructured and structured overlay networks. We outline the respective algorithmic aspects and we categorize the surveyed material so that pros and cons of each method can become apparent in their context.

We employ both tabular and pictorial means to articulate features of the proposed methodologies and outline both strengths and weaknesses. To ascertain the effectiveness of each proposal, we predominantly use three criteria that capture the behavior of the algorithms discussed: *efficiency*, *overhead* and *scalability*. For each of these criteria, we use a coarse-grained, three-level, gradation to denote how well the algorithms do, namely, *low*, *medium* and *high*. Each criterion has its own interpretation of the above literal values.

We use the term *efficiency* to gauge the competence of a given algorithm to match the overlay to the underlying network. Minimization of duplicate messages traveling the network, quick object location and preservation of search scope as well as ability and effectiveness to adapt in high churn environments are some of the characteristics we consider when grading an approach on this criterion.

With *overhead*, we, coarsely, quantify the additional cost a given approach imposes to the underlying resources when applied to (i) create, (ii) adapt, and (iii) maintain the overlay network while preserving its topology awareness. Computing and communicating peer join and departure, message forwarding, updating routing tables are some of things that we take into account to perform our assessment. Note that this is the one criterion of the three we use that its literal values "low" and "high" must be interpreted as "better" and "worse" respectively.

For the *scalability* criterion, we consider system behavior with respect to the number of participating peers it can reliably support. We characterize protocols as being lowly, mediumly or highly scalable if they are proven to effectively support hundreds, tenths of thousands, or millions of nodes, respectively. One important consideration here is the extent of self-organization of a system's peers and thus the overall distributed nature of the overlay network. The least a network relies on global knowledge to perform it functions the smoother it scales.

3. Unstructured P2P networks

In this section, we present algorithms that tackle the topology mismatch problem in unstructured P2P networks. We classify them based on their use of the overlay structure, their message forwarding scheme for peer communication and the techniques they use for detecting proximity.

3.1. Algorithms for unstructured architectures

To improve over *Gnutella*'s "blind flooding" approach, Yang and Garcia-Molina (2002) proposed a practical and easy to implement solution weaved around three different message forwarding methods, namely *iterative deepening*, *directed BFS* and *local indices*.

In **iterative deepening**, the search is performed on a BFS tree with multiple preset depths. The depth limit is iteratively increased by the source node for each query, based on the quality of the results. The source node may issue a new request with increased depth limit that will trigger the nodes at the last visited depth level to resume the

347 search thus avoiding restart of the entire search process and consequently reducing the load on the nodes of the upper levels of the tree.
 348 Its major drawback is the delay between successive iterations, as the source-node has to examine the results with each attempt before deciding to either quit or “unfreeze” the query.
 349

350 The **directed BFS** tries to avoid the aforementioned delay by forwarding query messages only to a selected subset of available neighbors. The selection criteria varies, from the number of results received or the distance (hops) traveled to locate these results, to the bandwidth, or the query load of the neighbor. That way, fewer nodes are visited but quality of query responses is maintained, to a large degree, on the premise that the selected heuristics can direct the search to the right path.
 351

352 In **local indices**, each node indexes data hosted by neighboring nodes within a radius of r hops and uses this local index to answer queries on behalf of other nodes. That way, local indices, greatly reduce the aggregate bandwidth usage of the network and improve query efficiency. However, updating such indices in the presence of frequent node joins/departures may introduce significant overhead should the radius is kept broad.
 353

354 Below, we outline how the techniques in [Yang and Garcia-Molina \(2002\)](#) fare against the three major performance criteria of [Section 2](#).
 355

	Efficiency	Overhead	Scalability
Iterative deepening	Low	Low	Medium
Directed BFS	Medium	Medium	Medium
Local indices	High	Medium	Medium

356 A *delay-aware P2P* system termed **DAPS** was introduced in [Zhang and Lin \(2005\)](#). DAPS seeks to attain reduced look-up times by dividing peer routing tables into several sectors of increasing delay. The source node that issues the query designates the delay boundary it may tolerate, by providing a *pruning factor*. In this context, user requests are forwarded only to nodes whose expected delay is less than or equal to the indicated boundary. DAPS primarily focuses on user “experience” and deploys an end-to-end delay monitoring mechanism that may enable the clustering of routing tables. In a dynamic environment, the formation of such accurate routing tables may be an elusive goal. In terms of the three performance criteria, DAPS stands as follows:
 357

	Efficiency	Overhead	Scalability
DAPS	Low	Low	Medium

358 The key objective of the **Gia** system ([Chawathe et al., 2003](#)) is to help alleviate the scalability omnipresent in unstructured P2P file-sharing systems. At first, Gia replaced Gnutella’s blind flooding with random walks. Although this adoption was a step in the right direction ([Lv et al., 2002](#)), issuing a single copy of the query within the network effectively reduces the search scope and may negatively affect the success rate of the query in question. To overcome this limitation, Gia introduces a token-based flow control mechanism that gradually redirects queries to nodes which are more likely to answer. This flow control mechanism also helps prevent node overloading as each peer “announces” the number of query requests it can handle, in terms of tokens, to its neighbors; to this end, peers only forward query requests to nodes that they previously received tokens from. Further refinement to the search mechanism is the support for *one-hop replication* of pointers to content. Gia also acknowledges the heterogeneity in peer bandwidth, processing power, disk speed, etc., of P2P nodes and uses this information when connecting nodes to each other. By using a topology adaptation algorithm, Gia places low capacity nodes within short proximity to peers with high performance features. This topology adaptation algorithm is based on the metric each node maintains about its satisfaction – ranging between 0 and 1 – for the neighbors with which it finds itself associated. Through message exchanges, a peer can establish new connections or drop superseded ones to improve its satisfaction. Despite the fact that Gia’s

405 topology adaptation algorithm improves the network’s scalability, it falls short in addressing the mismatch problem as considerations for the underlying network are not handled in explicit terms ([Liu et al., 2005b](#)). Moreover it reduces the search scope, exhibits poor performance in the worst case when matching data is not found quickly ([Pyun and Reeves, 2004; Cai and Wang, 2004](#)) and can potentially build disconnected topologies ([Srivatsa et al., 2006](#)).
 406

407 In terms of the three stated criteria, Gia’s approach is as follows:
 408

	Efficiency	Overhead	Scalability
Gia	Medium	Low	Medium

409 **Location-aware Topology Matching (LTM)** ([Liu et al., 2004a](#))
 410 seeks to optimize an overlay P2P structure based on the physical
 411 topology. To achieve this, peers issue special messages called *TTL-detectors*
 412 whose *TTL* values are 0 or 1; in this regard, peers discover
 413 1- and 2-hop neighbor sets, designated as N and N^2 respectively, and
 414 proceed to compute communication costs. Time-stamps are used to
 415 derive network latency measurements that are then used to improve
 416 the overlay network without sacrificing the search scope. Each node
 417 compares the latency information received from its direct neighbors;
 418 peers with longer latencies are placed on a *will-cut* list where they
 419 remain for a certain period of time after which they are finally elimi-
 420 nated and are placed on the peer’s *cut-list*. Thus, low-productivity
 421 connections are dropped and replaced by more efficient ones, reduc-
 422 ing the latency on the overall network. Although LTM improves the
 423 overall efficiency of the P2P network, it is unable to offer any war-
 424 ranty for effectively addressing the mismatch problem. LTM fares as
 425 follows regarding our three criteria:
 426

	Efficiency	Overhead	Scalability
LTM	Medium	Medium	Medium

427 Another problem with topology unaware systems is the dupli-
 428 cation of messages due to cycles in the network graph. Interestingly,
 429 such cycles appear even along the correct forwarding path. [Zhu et al. \(2008\)](#)
 430 focuses on this exact deficiency of overlay networks and intro-
 431 duces the **Distributed Cycle Minimization Protocol (DCMP)**, a dy-
 432 namic, fully distributed method that removes cycles; this is accom-
 433 plished without sacrificing overlay connectivity, resilience and other
 434 key properties of unstructured P2P architectures and by avoiding a
 435 hierarchical organization of peers. Once a cycle is detected in DCMP,
 436 the most powerful node in that cycle is elected as the *Gate Peer* and
 437 the cycle is then broken in that place that it will result in the mini-
 438 mization of the distance between the *Gate Peer* and all other nodes
 439 that are currently part of the cycle. This process is managed by using
 440 two specialized message types namely, *Information Collection Mes-
 441 sage (ICM)* and *Cut Message (CM)*. DCMP bases its operation on mes-
 442 sages whose travel is limited by an imposed *TTL* value (max set to 7
 443 for most cases). This inherently limits the protocol in detecting cy-
 444 les that span for more than *TTL* nodes. Even though cycle elimina-
 445 tion does improve network performance, it cannot directly contribute
 446 to the solution of the topology mismatch problem, while its perfor-
 447 mance in high-traffic and churn environments and its ability to re-
 448 tain important characteristics of a Gnutella-like topology (i.e., degree
 449 distribution, average peer distance, diameter e.t.c.) has been doubted
 450 ([Chandra et al., 2010](#)). On the other hand, overhead is greatly reduced
 451 due to its “lazy” broadcasting of control messages, as opposed to, for
 452 example, LTM’s periodic approach ([Zhu et al., 2008](#)). The expected
 453 DCMP behavior as far as our three criteria is:
 454

	Efficiency	Overhead	Scalability
DCMP	Medium	Low	High

455 In [Cohen and Shenker \(2002\)](#) and [Lv et al. \(2002\)](#), replication is
 456 used as a way to improve inefficient blind search. The observation is
 457 that as the number of replicas increases in the network, it would be
 458 relatively easy to locate items even if search is random. Concretely,
 459

461 three three approaches for replication allocation are proposed: ***uniform***,
 462 ***proportional***, and ***square-root replication***.

463 In the *uniform model*, all objects are replicated without taking into
 464 account the query distribution, while in the *proportional*, objects are
 465 replicated analogously to their query rate, so that frequently queried
 466 items can be found more often. Experimental outcomes indicate that
 467 the uniform allocation minimizes the maximum search length and
 468 so it can reduce the time spent on unsuccessful searches. The pro-
 469 portional strategy effectively decreases the search time for popular
 470 items, but suffers when needing to locate rare items. For allocation
 471 ratio “between” 1 and the ratio of the query rates, the authors prove
 472 that the average performance on successful queries is always better
 473 than in uniform and proportional models. Finally, with the *square-*
 474 *root model* they attempt to find the golden ratio in “between”, by sug-
 475 gesting to set the number of object replicas to the square root of the
 476 searching rate for an object.² Theory in Cohen and Shenker (2002)
 477 suggest that square-root minimizes the expected search size on suc-
 478 cessful queries, a claim that is supported by simulations (Lv et al.,
 479 2002).

480 Lv et al. (2002), present a *k-walker* query algorithm and show that
 481 using it along side square-root allocation results in performance simi-
 482 lar to Gnutella’s query flooding method but incurs up to two orders
 483 of magnitude less network traffic. Also checks three strategies for
 484 replica placement, namely *owner replication* where an object is repli-
 485 cated at the requesting node, *path replication* that caches the object
 486 along the path from the requesting to the providing node and *random*
 487 *replication* where the replication is taking place randomly on visited
 488 nodes. Observations in this work are not directly applicable to struc-
 489 tured DHTs, because it is assumed that the lookup time for an object
 490 depends only on the number of replicas and not the placement strat-
 491 egy (Ramasubramanian and Sire, 2004).

492 The table below outlines how the three allocation approaches
 493 fare:

	Efficiency	Overhead	Scalability
Uniform replication	Low	Low	Medium
Proportional replication	Low	Medium	Low
Square-root replication	Medium	Medium	Medium

494 **Narada** (Chu et al., 2000; 2001; 2002) is a generic protocol for
 495 creating self-adapting overlay networks capable of application-layer
 496 multicast communications without requiring IP multicast infrastruc-
 497 ture at the network layer. Although IP-multicast would present a
 498 choice in implementing *Narada*, it is in general considered that it vi-
 499 olates the stateless design of the current Internet. Despite the fact
 500 the *Narada* was not designed as P2P system per se, it was the first
 501 (along with Scattercast (Chawathe, 2000)) to consider the feasibility
 502 of overlay-based, application-layer services over the Internet while
 503 taking into account bandwidth and latency properties of the under-
 504 lying physical infrastructure. To overcome the inefficiency caused by
 505 topology mismatch, *Narada* builds a highly connected graph, termed
 506 *mesh*, with each source featuring its own minimum spanning tree. A
 507 gossip-protocol was deployed for the creation of these spanning trees
 508 (Xing-feng et al., 2008). The graph and trees are dynamically updated
 509 as nodes keep joining or departing the network. The protocol aims
 510 to ease the physical link stress, the overall resource usage as well as
 511 the relative delay among end-systems. Unfortunately, the main limi-
 512 tation for *Narada* is that although it works reasonably well for small
 513 groups, it does not scale well for larger networks. Hence, it is not a
 514 suitable choice for potentially large P2P networks:

	Efficiency	Overhead	Scalability
Narada	Medium	High	Low

Along with *Narada*, the ***Adaptive Overlay Topology Optimization (AOTO)*** (Liu et al., 2003) is one of the first attempts to address the topology mismatch problem. *AOTO* is a distributed algorithm that seeks to optimize the usage of the underlying physical resources and operates in two phases: *Selective Flooding* and *Active Topology*. In *Selective Flooding*, a minimum spanning tree is built for each peer and its immediate neighbors so that queries do not flood the entire network while preventing their search scope from shrinking. This way some neighbors become non-flooding. During *Active Topology*, each peer replaces independently such non-flooding neighbors, with closer nodes as an attempt to revise the overlay links so that they can reflect more closely the physical network topology. Picking a replacement out of these non-flooding neighbors follows a random policy called *Randomized AT* algorithm. To accomplish the above actions, a peer has to keep track of its communication costs with all its neighbors (e.g., network delays) as well as the costs between any pair of neighbors. The *randomized AT* algorithm is applied by a source peer every time its neighbor list is updated or an updated neighbor cost table is received. *AOTO*’s performance regarding the three criteria is as follows:

	Efficiency	Overhead	Scalability
AOTO	Medium	High	Medium

The ***Adaptive Connection Establishment (ACE)*** approach (Liu et al., 2004b) uses the network delay as a metric to estimate the costs between nodes. Each peer computes the costs to its logical one-hop neighbors and forms a *neighbor cost table (NCT)* using a special routing message type. Any pair of neighboring peers exchange their *NCTs* and so a minimal overlay topology can be formed. Based on the obtained *NCTs* a minimum spanning tree for each peer and its one-hop neighbors is created. Finally, neighbors located physically far away are replaced by physically-closer counterparts. In particular, a peer *S* iteratively probes the distance *d* between itself and each of its non-flooding neighbor nodes *G* as well as the distance between *S* and *G*’s neighbors designated as *H*. If $d_{SG} > d_{SH}$, then link *SG* is replaced by link *SH*. If, on the other hand, $d_{SG} < d_{SH}$ then we have the following options. Either $d_{SH} < d_{GH}$ in which case link *SH* is kept or $d_{SH} > d_{GH}$ in which case *S* will keep probing other *G*’s direct neighbors. The above optimization is conducted within 1-neighbor closure using as base, a peer and checking all its direct neighbors. Evidently the scope of such an operation could be extended. Should a larger scope be used, a better topology matching can be obtained at a greater computational overhead.

	Efficiency	Overhead	Scalability
ACE	Medium	Medium	Medium

The ***Scalable Bipartite Overlay (SBO)*** (Liu et al., 2007) reduces the overhead of creating and maintaining a minimum spanning tree by randomly dividing the nodes into two groups –*reds* and *whites*– and assigning different tasks to the two groups. When a peer joins the network, it is randomly assigned with an initial color (red or white). Then the network bootstrap host node furnishes the joining peer with a list of active peers along with their color information. The joining peer uses this list to establish connections to differently colored peers. In this regard, all peers form a bipartite overlay. Using the network delay as metric, white-peers measure distances from red counterparts and report the encountered red neighbors. Having information on all their 2-hop neighbors (N^2) red-peers create minimum spanning trees for the neighbors in question and assign efficient 2-hop forwarding paths. This process can render a white-peer *E* non-forwarding neighbor of the red-peer *P*. Direct neighbor replacement is a process conducted by *E* to replace *P* with a $P' \in N^2(P)$ as its new neighbor. This adaptation tackles the topology mismatch problem by reducing message duplication and shorten response times caused by the problem identified as *Revisit Not known (RN)*. RN is the situation when a node is visited some times as a relay stop before it is visited

² An interesting application of the square-root principle has also been explored in the context of topology reconstruction, by ensuring node degree to be proportional to the square-root of its content popularity (Cooper, 2005).

575 as an overlay peer, thus creating duplicate unnecessary messages to
 576 the network and delaying query responses. Even though this attempt
 577 is based on a simple and elegant solution it is reported to not guarantee
 578 performance in terms of average communication delay or search
 579 scope (Hsiao et al., 2009). In regards to the three stated criteria, SBO
 580 behaves as follows:

	Efficiency	Overhead	Scalability
SBO	High	Medium	Medium

581 The distributed heuristic termed **Two-Hop-Away Neighbor Comparison and Selection (THANCS)** (Liu et al., 2005a; Liu, 2008) attempts to minimize overlay hop costs. THANCS is essentially a *local search method* as it aims to find a locally optimum solution by exploiting knowledge within a 2-hop radius. The algorithm consists of two main components: *piggybacking neighbor distance on queries* and *neighbor comparison and selection*.

582 The *piggybacking component* requires peers to probe immediate neighbors using delay distance measurements and store this information locally. The algorithm introduces a special query message type, the *Piggy Message (PM)* which includes information about the neighbor identification (*IP*) and measured distance (*d*). PMs are piggybacked to normal *Gnutella* messages. When a node *P* receives a query from *Q*, it constructs a *PM_{PQ}*, piggybacks it to the query and forwards it to *P*'s neighbors. As soon as a neighbor detaches a received *PM_{PQ}*, records the distance *d_{PQ}* and processes the query. Selection of which incoming queries should be piggybacked with a *PM* is determined using either a *pure probability-based (PPB)* or a *new neighbor triggered (NNT)* policy.

583 The *neighbor comparison and selection component* defines that a peer *S* probes all his 2-hops away neighbors (a set denoted as *N²(S)*) not probed thus far. Let *P* be a 1-hop distance neighbor of *S*, and *Q* be a probed peer by *S*. When *S* receives a message piggybacked with message *PM_{PQ}*, the following 2 cases are identified:

- 605 1. If *Q* is already a direct neighbor of *S* then we check the distances *d_{SQ}*, *d_{SP}* and *d_{PQ}*. If either *d_{SQ}* or *d_{SP}* is the longest, then the longest link will be placed in a *will-cut* list.³ If *d_{PQ}* is the longest, then nothing is done by *S*; being fully distributed, neighbor comparison and selection process, will have the opportunity to handle *PQ* link when initiated by peers *P* or *Q*.
- 611 2. If *Q* is strictly a 2-hop neighbor of *S* and have never probed *Q* in the past, stores distance *d_{SQ}* and checks distances *d_{SQ}*, *d_{SP}* and *d_{PQ}*. If *d_{SQ}* is the longest, *S* will not create link *SQ*. If *d_{SP}* is the longest, *SQ* will be created and *SP* will be put into the *will-cut* list. If *PQ* is the longest, links *SP* and *SQ* will be preserved expecting that *P* or *Q* will disconnect *PQ* later.

617 As discussed above, THANCS is fully distributed and needs only minimum knowledge of, at most, a 2-hop distant peers. This features 618 renders the method a good candidate for large overlay networks. Also 619 its topology adaptation helps construct a well fit overlay with respect to the underlying network—at least with respect to network 620 distances. In addition, will-cut list and distance caches (which store 621 already probed peers) minimize the unnecessary messages for the 622 network maintenance. THANCS reportedly improves broadcast performance 623 the extent of which, though, depends on the underlying 624 network since THANCS does not provide the performance guarantee 625 for the improvement of any given overlay topology (Hsiao et al., 626 2010).

	Efficiency	Overhead	Scalability
THANCS	Medium	Low	High

³ A peer will not send or forward queries to connections in its will-cut list but will preserve them for some time in order not to affect ongoing response messages traveling the inverse path.

The **Hops Adaptive Neighbor Discovery (HAND)** algorithm (Chen et al., 2006) uses a fully-distributed triple-hop adjustment strategy, applied to a network graph *G* to create an optimal graph *G**, free of the implications of topology mismatch. This optimality is attained if all peer hop sequences (v_1, v_2, \dots, v_k) of *G* exist in *G** and are in the same order. The latter indicates that in practice triple sequences (v_1, v_2, v_3) are used. A mismatch is detected as follows: should we want to verify a sequence, say $v_2 - v_1 - v_3$, two probing messages are dispatched from v_1 to v_2 and v_3 and yield delays of x and z respectively. When the probing message arrives at v_2 , it gets forwarded directly to v_3 . Similarly, the message reaching v_3 is directly forwarded to v_2 . The above two forwarding actions seek to quantify the corresponding delays of (v_2, v_3) and (v_3, v_2) for the physical path *y*. If $y = z - x \pm \epsilon$, the sequence $v_2 - v_1 - v_3$ is mismatched and should be adjusted to $v_1 - v_2 - v_3$ by deleting edge (v_1, v_3) and then adding a new (v_2, v_3) . If $y = x - z \pm \epsilon$, sequence $v_2 - v_1 - v_3$ is mismatched and should be adjusted to $v_1 - v_3 - v_2$ by deleting edge (v_1, v_2) and adding a new (v_3, v_2) . The ϵ is a small positive real number denoting additional delays caused by possible forwarding and jitter delays. The advantages of HAND are that it (i) does not need any clock synchronization, (ii) is a fully distributed algorithm, (iii) involves low traffic overhead, (iv) can be used in dynamic P2P environments, and (v) furnishes low query response times. In terms of the three stated criteria, HAND fares as follows:

	Efficiency	Overhead	Scalability
HAND	Medium	Low	High

The **Adaptive Peer Selection (APS)** approach uses machine learning techniques to form peer selection strategies based on past experience (Bernstein et al., 2003). A decision tree is used to rate peers based on information collected for the characteristics of established connections. Such features include connection link load, bandwidth, and past uploading experience. Subsequently, a Markov decision process is used to shape the policy for switching among the peers. The success of this approach depends on how fast the introduced peer selection functions. Admittedly, APS is slow due its on-line learning process and inherent complexity (Zeng et al., 2009). APS's behavior regarding the three criteria is as follows:

	Efficiency	Overhead	Scalability
APS	High	High	Low

The **Innocuous Topology Aware Overlay Construction (ITA)** approach (Papadakis et al., 2013) seeks to offer both an overlay formation approach and an effective way to carry out searching. When constructing the overlay, ITA exploits the notion of *short* and *long* connections. Should *N* be the number of network peers, $\alpha \leq 1$ a system-wide magic number and *x* an α -related latency threshold, the bootstrapping peer randomly selects $\alpha^* N$ "close" (latency below *x*) and $(1 - \alpha) * N$ distant (latency above *x*) nodes as its neighbors. Of course, the bootstrapping peer is not able to check all the peers in the network to find the global best for the above sets. The authors prove that latency measurements against $30/\alpha$ randomly selected peers result to a good estimation. Searching occurs in two phases: first, the querying node floods its distant neighborhood with *TTL* = 1. Subsequently, peers receiving a query over a "long link" commence a local flood with *TTL* = *ttl* where *ttl* is system-defined. The main objective of the overlay construction phase is to yield a network that exhibits low clustering. In turn, this is a beneficial characteristic for random graphs as it can lead to a larger coverage of peers and at the same time help reduce duplication of messages traveling around the network. This combination offers efficient lookups that pose minimal or no negative impact to other mechanisms possibly employed by the P2P systems (e.g., 1-hop replication or dynamic querying mechanisms). Experimental results in the same paper suggest a 50% reduction in communication latency among peers by cutting off up to 25% of the

688 IP-level traffic generated. Regarding the three criteria, *ITA* is placed as
 689 follows:

	Efficiency	Overhead	Scalability
ITA	Medium	Low	Medium

690 **EGOIST** (Smaragdakis et al., 2008) is a set of algorithms used to
 691 construct and manage overlay networks. *EGOIST* utilizes a selfish ap-
 692 proach in the sense that every participating peer continuously up-
 693 dates its neighbors so as to minimize the sum of distances to all desti-
 694 nations under shortest-path routing. A newly arriving peer, randomly
 695 connects to an already participating node through a bootstrap server.
 696 Once connected, the peer starts receiving information via a link-state
 697 mechanism and thus, after some time, it has a complete picture of the
 698 overlay graph. Then, the peer estimates the delay to all other nodes
 699 to determine its potential neighbors and ultimately connects to the
 700 overlay using some policy; such a policy might be for example the
 701 minimization of the average delay to all its neighbors. *EGOIST* requires
 702 extensive resource usage to continuously update the overlay connec-
 703 tions for all nodes in the system. This needs $O(n^2)$ measurements.
 704 Fortunately, each node does not need to do these measurements for
 705 maintenance (monitoring and updating) with all other participating
 706 nodes, but just with a number of $k < n$ nodes that it chooses to es-
 707 tablish links with. The latter yields a reduced $O(kn)$ time complexity
 708 for maintenance while complexity $O(n^2)$ is required less frequently
 709 and only when nodes re-evaluate their connections. *EGOIST* fares as
 710 follows:

	Efficiency	Overhead	Scalability
EGOIST	High	High	Low

711 The main objective of the **Biased Neighbor Selection (BNS)**
 712 approach (Bindal et al., 2006) is to strengthen BitTorrent's (Cohen,
 713 2003) locality by carefully choosing most of the neighbors of a peer
 714 to come from the same ISP, while simultaneously adhering to the
 715 near-optimal download performance of the protocol. By and large,
 716 BitTorrent deploys mechanisms that have proved aggressive to ISP
 717 networking and accounting as they function in a "without-borders"
 718 approach. The default BitTorrent specification designates 35 connec-
 719 tions for each peer regardless if such connections are placed within
 720 or outside the borders of the ISP that hosts the node. *BNS* proposes
 721 that only k out of these 35 connections be established with nodes
 722 beyond the ISP-borders; the k connections offer an extended and
 723 "global" view of the network at large and seek to strike a balance
 724 between the load imposed inside and outside the ISP. The latter re-
 725 frains peers from exchanging traffic through expensive network links
 726 should there exist alternative local connections offering the required
 727 services. *BNS* is realized through (1) the modification of the tracker
 728 so that ISP-local resources are rapidly identified (2) the use of P2P-
 729 traffic-shaping devices on ISP edge-routers. In Bindal et al. (2006),
 730 a combination of biased neighbor selection along with bandwidth
 731 throttling and caching techniques for near-optimal lookup results
 732 is proposed. The effectiveness of *BNS* is debated in Ren et al. (2010)
 733 where experimental results suggest that most peers in a swarm
 734 reside in different autonomous systems (ASs). Consequently, clients
 735 frequently do not have enough peers to leverage intra-AS connections
 736 (Ren et al., 2010). Moreover, *BNS* deployment requires knowledge of
 737 ISP network mappings and awareness of possible changes that occur
 738 in the infrastructure limiting its widespread adoption; the following
 739 table qualitatively depicts the expected behavior of *BNS*:

	Efficiency	Overhead	Scalability
BNS	High	Low	Medium

740 **Ono** (Choffnes and Bustamante, 2008) is a protocol that helps con-
 741 tain BitTorrent traffic within individual ISPs and enhances the down-
 742 loading rates by favoring connections within the borders of a single
 743 autonomous system. Contrary to *BNS* (Bindal et al., 2006), *Ono* leads
 744 to improved performance without requiring any explicit cooperation

between ISP subscribers or any additional infrastructure and network
 topology information. *Ono*'s selection approach is landmark-based
 and leverages existing CDN-infrastructure for peer distance estima-
 tion. CDNs already use both static (i.e., geographical, IP-clustering)
 and dynamic (i.e., network delay measurement) information for their
 replica selection. Thus, the algorithm leverages the observation that
 peers that are redirected to a specific CDN replica does not only mean
 that these are close by this replica but to each other as well. The
 redirection behavior is modeled in terms of *ratio map*, a vector of $<$
replica-server, ratio $>$ tuples, where *ratio* is the percentage of times
 the CDN redirects the peer to the specific *replica-server*. The proto-
 col's bootstrapping phase consists of the incoming peer performing
 DNS lookups to CDN-names to build its redirection information. Ini-
 tially, the DNS polling interval is set to 30 s and increases by 1 min ev-
 ery time redirection information to the CDN remains unchanged after
 the lookup. The DNS polling interval is halved if pertinent redirection
 information has changed. To avoid the bootstrapping phase when the
ratio map is sufficiently fresh, *Ono* persists the map after the end of a
 BitTorrent session. Experimentation with *Ono* yields minimization of
 inter-ISP traffic costs and achieves near optimal performance of Bit-
 Torrent download (Bindal et al., 2006). The ISP-friendly design of *Ono*
 however has been questioned when it comes to Internet-scale (Piatek
 et al., 2009; Liu et al., 2009; Cuevas et al., 2009); its qualitative rating
 stands as follows:

	Efficiency	Overhead	Scalability
Ono	Medium	Medium	Low

In Liu et al. (2009), three techniques that "inject" locality awareness in BitTorrent-like applications are discussed. The first macroscopic-level technique focuses on improving the neighbor se-
 769 lection process. When a peer asks a tracker to join, the later sorts
 770 its swarm peers according to their hop-count distance from the re-
 771 questing peer and sends out the top- k list of nodes (e.g., $k = 50$).
 772 The second technique, applied in an intermediate-level, manipulates
 773 BitTorrent's chocking/unchocking mechanism. A peer unchoke its 4
 774 closest, in terms of autonomous system hop-count, interested neigh-
 775 bors. The same applies also when the peer turns to the seeding state;
 776 this intermediate-level approach favors least distance and is in con-
 777 trast to the original BitTorrent implementation which favors upload-
 778 speed. The last technique operates at a microscopic-level by choos-
 779 ing to pick the next chunk to download based on a locality-first pol-
 780 icy; this picks first the chunk that is closer, as opposed to BitTorrent's
 781 rare-first chunk picking policy. The distance value of a chunk is com-
 782 puted as the mean value of the distances of the peers that posses
 783 it. Through experimentation on a real-world Internet topology simu-
 784 lated in PlanetLab, the locality-based approach achieves traffic opti-
 785 mization. Nevertheless, it does not do that well when compared with
 786 the random approach of the standard BitTorrent protocol.

All three techniques in Liu et al. (2009) seek to harness the locality
 790 of the basic functions of the BitTorrent protocol. Even though they do
 791 not restructure the network to map better to the underlying infras-
 792 tructure, they favor the intra-AS over inter-AS communications thus,
 793 reducing the overall communication cost. Again, there is a trade-off
 794 between reducing inter-domain traffic and fairness among peers in
 795 terms of the data the peers upload (Liu et al., 2009); as far as the
 796 three criteria, the overall approach fares as follows:

	Efficiency	Overhead	Scalability
Liu et al. (2009)	Medium	Low	Medium

The key objective of **TopBT** (Ren et al., 2010) is to enhance proxim-
 798 ity awareness in BitTorrent without the need of additional infrastruc-
 799 ture. *TopBT* is built on the conjecture that a good peer selection metric
 800 should take into account both the downloading speed and the net-
 801 work topology. The metric proposed is defined as the ratio of down-
 802 load to upload rate d/u divided by either the link-level l or the number
 803 of routing hops required a . This metric should guarantee selection of
 804

peers with high download rate, low reciprocal upload demand and decreased routing hops. The above *TopBT*-metric is applied in various aspects of the original *BitTorrent* protocol to "inject" topology awareness for improved peer selection. More concretely, *TopBT* modifies: (1) the peer list the tracker returns when someone first tries to connect to the swarm, (2) the initial connection establishment, and (3) the unchoking mechanism of *BitTorrent*. *TopBT*'s key features are that it does not rely on any Internet infrastructures (ISPs or CDNs) whatsoever and it cooperates with already deployed *BitTorrent* clients. It also supports both link-hop and AS-hop metrics to identify proximity in different levels to reduce overall traffic while preserving downloading performance. Through experimentation on hundreds of *PlanetLab* and residential hosts, *TopBT* offers more than 25% traffic reduction and 15% faster downloads compared to popular *BitTorrent* implementations. In reference to the stated three criteria, *TopBT* fares as follows:

	Efficiency	Overhead	Scalability
<i>TopBT</i>	Medium	Low	High

The **Underlying Topology-Aware Peer Selection (UTAPS)** (Li et al., 2008) presents an enhanced peer selection strategy. *UTAPS* operates in two stages: in the first, it collects information to develop a picture of the underlying topology and in the second, it makes its peer selection based on this knowledge. In the first stage, *UTAPS* employs network tomography, a technique that probes from a large network's endpoints to infer its internal characteristics (Coates et al., 2002). Upon arrival of a new peer, the tracker trace-routes it to obtain some basic knowledge including IP address, routers involved, RTT and number of hops traversed. The more the peers in the swarm, the better picture a tracker has for the underlying infrastructure. The tracker provides this information to the newcomer in the form of the bootstrapping peer list. These newly joined peers traceroute the tracker peer list themselves and return back to the tracker to further enhance the tracker's view of the infrastructure. In the second stage, *UTAPS* employs a set of heuristics which target those peers that expose low RTT and are found within certain hop-counts away from the routers observed during the initial traceroute. Peers running *UTAPS* instead of the random peer selection, achieve better downloading rates and at the same time offer reduced burden on the underlying ISPs (Li et al., 2008). *UTAPS* stands as follows in regard to the three criteria:

	Efficiency	Overhead	Scalability
<i>UTAPS</i>	Medium	Medium	Medium

A clustering approach that differentiates peers encountered in a swarm in those that are local, intra-ISP and inter-ISP is presented by Qin et al. (2009). The classification is carried with the help of the core routers used while peers communicate with the tracker. It is assumed that even though the network may be highly volatile, these core routers are usually more stable. For this, a newly arriving *BitTorrent* peer, trace-routes the tracker and stores vectors containing information like the IP address and the hop number of a traced router, as well as the link delay of the traced router and the previous hop along the trace-route path. These vectors are reported to the tracker which uses this information to classify the router through a k -means classification algorithm, with $k = 3$. For peer selection, Qin et al. (2009) proposes a biased approach where the tracker returns c close peers and $d = N - c$ distant peers, where N is the length of the returned list. In choosing c , the tracker employs an iterative search process first in the peer's local neighborhood. If there is not a sufficient number of peers available, the tracker goes to intra-ISP and subsequently, if needed, to the inter-ISP cluster. Experimentation in a controlled environment indicates up to 5% faster downloading times as well as up to 22% reduction of the total cross-ISP traffic. This clustering approach fares as follows with regards to the three criteria:

	Efficiency	Overhead	Scalability
<i>Qin et al. (2009)</i>	Medium	Medium	Medium

In Qiu et al. (2007), two **Peer-exchange Routing Optimization Protocols (PROPs)** are introduced to help adjust the neighborhood graph of an overlay network while reducing the network's overall link latency. PROPs are weaved around the concept of "exchange" of neighbors among peers; this is carried out so that participant node can collectively benefit from the attained reduction in network delays. This "collaboration", is what differentiates this approach from others by allowing two peers to optimize their neighborhood environment, in contrast with simply allowing each node to "selfishly" choose its own strategy. The PROPs follow two phases: during the *warm-up*, a node u probes its neighbors to collect distance information. Subsequently, u contacts, at a fixed time rate *timer*, a random node v , n -hops away. New distance information is calculated independently by nodes u and v , now for the hypothetical state when the potential exchange would occur. If the benefit gained, in terms of reducing the average distance among the peers involved, is above some predefined threshold then the exchange does occur. Otherwise, no further action is taken. The second *maintenance* phase differs from the warm-up in that the probability of peers to be probed and the *timer* interval that this happens now depends on past peer exchange results.

In the generic form of the *PROP-G* protocol, a peer can exchange all of its neighbors with those of another peer. This essentially results in the two nodes exchanging positions so the topology and connectivity of the overlay is not affected by the operation of the algorithm. The optimized version *PROP-O* permits peers to exchange selected sets of their neighbors. It is not allowed to exchange neighbors along the path of the nodes that perform the exchange to ensure that in the end of the process the peers will remain connected. Additionally, the algorithm needs to preserve the characteristics of the network (i.e., the graph remains connected and the degree of the nodes is unchanged), so the exchange always involves equal number of neighbors. The following table depicts the PROPs behavior regarding the stated criteria:

	Efficiency	Overhead	Scalability
<i>PROP-G</i>	Medium	Low	Medium
<i>PROP-O</i>	Medium	Medium	Medium

The **Distributed Domain Name Order (DDNO)** approach (Zeinalipour-Yazti and Kalogeraki, 2006) uses domain names to detect topologically-close nodes. The fundamental assumption of the approach is the nodes found in the same domain are also topologically close. *DDNO*'s outcome is a flat overlay topology which, with some adjustments, can be utilized in super-peer architectures as well. According to the algorithm, half of the possible connections of a node are used to connect to local peers (called *sibling* connections) and rest are used to randomly connect to the peers anywhere on the network (called *random* connections). The former ensure the reduction of long distance traveling for messages, while the latter secure the connectivity of the structure and prevent network partitioning.

Upon arrival, a peer n asks a *host-cache* to establish its random connections. To identify its siblings, n multicasts a specialized message l to locate the right candidate(s). Initially, the behavior of this message, is modeled as a random walker, until l reaches some peer d capable to guide it through its next step. This capability is enabled through a *ZoneCache*, a data structure that contains information for nodes accessible in an r -hop radius from d . Ultimately l will reach a node m that is candidate to become n 's sibling. Then, m issues a broadcast message to its own siblings and then all in the group will inform the initiating peer n of who is willing to accept new connections. The population as well as the maintenance of *ZoneCaches* across the network are key concerns when it comes to the efficient construction of the overlay. While seeking to address the mismatch problem, *DDNO* is a heuristic approach that uses *Split-Hash* and *dnMatch* algorithms to effectively encode domain names and determine domain

923 locality respectively. In terms of our three criteria, *DDNO* fares are
924 follows:

	Efficiency	Overhead	Scalability
<i>DDNO</i>	Medium	Low	Medium

925 The **Critical Topology-Aware Grouping (CTAG)** approach (Zhao
926 and Lu, 2006) advocates a grouping strategy of peers based on the *Internet
927 Assigned Numbers Authority (IANA)* and the respective *Regional
928 Internet Registry (RIR)*'s IP assignments. Based on these assignments,
929 nodes that belong in the same organization are always addressed
930 from the same block of IP addresses. CTAG exploits this observation to
931 construct topology-aware unstructured overlays. On top of this, *Adjacency
932 Measurement (AM)* is proposed, as a technique which uses the
933 longest matching IP segment criterion to compute node proximity.
934 CTAG focuses on both the construction of the overlay as well as its
935 constant and dynamic adaptation. In its first phase, called *bootstrapping
936 grouping*, the *Gnutella* web caching mechanism is modified so
937 that a newcomer may choose the closest, in terms of AM, cache to re-
938 trieve its bootstrapping list. Similarly, during its second stage, called
939 *dynamic revision*, the standard connection establishment mechanism
940 is also altered so that nodes with the highest AM metric are chosen.
941 When a node reaches its maximum number of neighbor connections,
942 it disconnects those neighbors with the lowest AM. CTAG behaves as
943 follows regarding the three criteria:

	Efficiency	Overhead	Scalability
<i>CTAG</i>	Low	Low	Medium

944 **Landmark Binning** (Ratnasamy et al., 2002a) partitions close-by
945 nodes into bins based on their distance from well known anchor
946 nodes across the Internet. To detect locality, peers mainly use net-
947 work latency (i.e., RTT). Despite the fact that such delays are not al-
948 ways accurate, they are used in Ratnasamy et al. (2002a) for they
949 are non-intrusive, transparent and easy to apply. For the binning to
950 work, a few anchor servers with known physical locations need to
951 be installed in strategic positions across the Internet. It is conjectured
952 (Ratnasamy et al., 2002a) that 12 such servers can prove suf-
953 ficient for the task. An arriving node measures its distance from these
954 landmarks and unilaterally decides to join a specific bin based on its
955 measurements. Specifically, the node measures its round-trip time to
956 each of the landmarks and orders the landmarks in decreasing order
957 of those RTT-values. Each permutation of the set of landmarks repre-
958 sents a specific bin. Should there be m landmarks adopted, $m!$ different
959 bins potentially exist. Peers that end up with the same ordering,
960 belong to the same bin, in the sense that if they experience similar
961 overall latencies from fixed network points; it is also very likely that
962 the peers in question are close to each other.

963 As the operation of the method is independent of the model incor-
964 porated by the overlay network, the approach can be applied with no
965 significant changes to either structured or unstructured P2P systems.
966 The rather obvious disadvantage of the approach lies in the landmark
967 servers that must be installed and maintained throughout the Inter-
968 net. Provided that a P2P-system often may have a few million nodes
969 connected at any given point in time, landmark servers do inherently
970 play a pivotal role in the operation of the network. The approach
971 has received a fair amount of criticism. For instance, Gummadi et al.
972 (2002) points out that end-hosts may bear the cost of latency mea-
973 surements and so, they may become bottlenecks. Castro et al. (2003);
974 Manku (2003); Zhang et al. (2004) argue that uneven distribution of
975 nodes ultimately lead to load imbalances and Ren et al. (2004) in-
976 dicates that its coarse-grained nature fails to effectively distinguish
977 among relatively close nodes. Finally, Xu et al. (2003) underscores the
978 potentially increased maintenance cost(s). The landmark binning be-
979 haves as follows:

	Efficiency	Overhead	Scalability
<i>Landmark binning</i>	Medium	Low	Low

The ***mOverlay*** (Zhang et al., 2004) approach addresses scalabil-
980 ity issues that might arise when static landmark servers are in use.
981 To this end, the use of dynamic landmarks is proposed. *mOverlay*'s
982 founding notion is that of a group that designates a set of peers found
983 in close proximity. This proximity is user-defined in the protocol and
984 may involve metrics including RTTs and network latencies. By and
985 large a clustering approach, *mOverlay* seeks to recreate small-world-
986 like properties by producing a two-level hierarchical structure: at the
987 top level, there are only connections among groups while at the bot-
988 tom, only *intra-group* connections occur among peers.
989

Clearly, identifying groups and accurately finding the closest
990 group to a peer is a fundamental concern in the creation of the over-
991 lay. Nodes are grouped based on their distance to the groups already
992 in the network, rendering the latter be the *dynamic landmarks* in the
993 process. For an incoming peer Q , the *grouping criterion* designates that
994 when the distance of Q and some group A 's neighbor groups is the
995 same as the distance between group A and group A 's neighbor groups,
996 then host Q should belong to group A . During network initialization
997 or when the grouping criterion is not met, new groups are created.
998 A peer may reach its group by expending at most $O(\log N)$ messages.
999 Finally, *mOverlay* maintains stability and constant overheads when a
1000 host either fails or departs the network; this is achieved through peri-
1001 odic cache updates and group leader selections, should a node either
1002 leaves or dies. In terms of the stated three criteria, *mOverlay* fares as
1003 follows:
1004

	Efficiency	Overhead	Scalability
<i>mOverlay</i>	Medium	Medium	Medium

3.2. Discussion on the algorithms for unstructured architectures

While surveying efforts to overcome the mismatch problem in
1005 the area of unstructured P2P systems, we came to identify four key
1006 methodologies utilized by the discussed approaches; they are:
1007

- (i) topology adaptation
- (ii) forwarding optimization
- (iii) caching and replication, and
- (iv) landmarking.

A key characteristic shared by many of the presented techniques
1013 is that they do not strictly adhere to a single methodology. As ap-
1014 proaches attempt to address the challenging problem of topology
1015 mismatch, they resort to heuristics and so, they do not often furnish
1016 "pure" approaches that exclusively belong to one of the four afore-
1017 mentioned methodologies. For instance, *mOverlay* combines topol-
1018 ogy adaptation with landmark binning while *Gia* weaves together
1019 three methodologies: topology adaptation, forwarding optimization
1020 and caching. The timeline in Fig. 12 further emphasizes this trend in
1021 the broader picture. All methodologies we discuss, have been iden-
1022 tified and put into practice, early on in the literature, while in sub-
1023 sequent years, the research community has mostly focused on fine-
1024 tuning the heuristic parameters of these methodologies in pursuit
1025 of the combination that performs best. This has led to better the-
1026 oretical understanding of the problem and its difficulty. Moreover,
1027 the combination of multiple methodologies has inevitably led to
1028 interesting feature/performance trade-offs that researchers need to
1029 tackle in real world applications. For example, the effectiveness of
1030 a caching/replication component can be undermined by a continu-
1031 ously adapting overlay that removes important links between peers.
1032 In particular, topology adaptation and forwarding optimization have
1033 been constant areas of focus of the research community. This is of
1034 no surprise given the volatile nature of unstructured P2P systems. If
1035 a technique is to be evaluated for a specific application domain, its
1036 overall approach as well as its underlying used methodologies have
1037 to be considered.
1038

Table 1

Summary table for unstructured algorithms.

Algorithm / Paper	Topology adaptation	Forwarding optimization	Caching/ Replication	Landmarking	Highlights	Pros/Cons
Yang and Garcia-Molina (2002)	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Iterative Deepening (ID). Directed BFS (DBFS). Local Indices (LI).	+ ID reduces the messages especially in upper levels of the tree. + LI reduces aggregate bandwidth usage and improves query efficiency. - ID needs evaluation time between iterations. - DBFS uses heuristics so it depends on their efficient choice. - LI add index update overhead which might be heavy process especially in high-churn systems.
DAPS - Zhang and Lin (2005)	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Clustered routing tables based on delay. Pruning flood, an iterative deepening and multiple BFS approach with a pruning boundary.	+ It is a hybrid system with both structured and unstructured characteristics.
Gia - Chawathe et al. (2003)	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Random Walks (RW).	+ RWs issue one copy of the query thus not flooding the whole network. - RWs can reduce search scope. + Compared to AOTO, ACE and SBO achieves faster convergence speed. - Creates more overhead than AOTO, ACE and SBO. - Needs synchronization of peer clocks.
LTM - Liu et al. (2004a)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	TTL detector (2-hop distance). Delayed low productive connection cutting.	- Does not consider shortcuts created by powerful peers when choosing to disable connections (only uses delay metric). + Drastically reduces duplicate messages. - Cannot detect cycles in distance bigger than the TTL value of the IC message.
DCMP - Zhu et al. (2008)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Cycle detection. Uniform replication.	+ Uniform replication reduces time spent on unsuccessful searches. + Reduces search time for frequent queries. k-walker querying scheme can reduce network traffic up to two orders of magnitude. - Proportional replication struggles in locating rare objects.
Cohen and Shenker (2002) & Lv et al. (2002)	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Proportional replication. Square root replication allocation. k-walker query scheme.	+ Mess and trees are kept up-to-date in high churn environments. - Works well only for small groups of peers.
Narada - Chu et al. (2000)	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Mess creation. Minimum spanning trees.	+ Spanning trees only to immediate neighbors so no flooding and at the same time no shrunk search scope. + Selective flooding effectiveness is detached from physical or overlay topologies. + The more logical neighbors, the more effective selective flooding becomes - High recalculation costs. - No sophisticated selection policy for candidate non-flooding peers.
AOTO - Liu et al. (2003)	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Minimum spanning trees. Peer proximity heuristic for removing costly links.	+ No flooding. + Less overhead compared to AOTO since computation is done within a certain diameter from the source peer. - Slow convergence speed. - Enhanced topology optimization comes at the expense of higher communication/computation overhead.
ACE - Liu et al. (2004b)	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Minimum spanning trees. 1-hop proximity heuristic.	+ Efficient in both static and dynamic environments. + Compared to AOTO incurs half the overhead. - Needs almost double the steps of LTM to converge (static or dynamic environments). + Completely distributed approach. + Presents trivial overhead compared to the query cost savings. + Convergent speed faster among AOTO, LTM, SBO.
SBO - Liu et al. (2007)	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Red/white bipartite overlay.	+ Does not shrink the search scope. - Design cannot be extended to support non-flooding-based systems.
THANCS - Liu et al. (2005a)	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Local optimum heuristic Piggybacking neighbor distance in queries	(continued on next page)

Table 1 (continued)

Algorithm / Paper	Topology adaptation	Forwarding optimization	Caching/ Replication	Landmarking	Highlights	Pros/Cons
HAND - Chen et al. (2006)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Triple-hop adjustment.	+ No need for clock sync. + Fully distributed. + Low overhead for the triple hop adjustment. + Applicable to both static and dynamic environments. + Low query response time. - Compared to LTM has lower traffic reduction and query response rates.
APS - Bernstein et al. (2003)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	machine learning adaptive mechanism.	+ Fully dynamic switching decision policy. - Low convergence due to the learning process.
ITA - Papadakis et al. (2013)	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Short/long connections. Local flooding.	+ Low clustering. + Large peer coverage. + Reduced duplication. + Low or no impact to other mechanisms of unstructured p2p networks (e.g. 1-hop replication, dynamic querying).
EGOIST - Smaragdakis et al. (2008)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Selfish shortest path routing	- constructs a global view of the network
BNS - Bindal et al. (2006)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	ISP clustering (tracker-side or ISP-side detection). Bandwidth throttling. Caching.	+ Localizes traffic within an ISP. + Preserves the efficiency of BitTorrent protocol. - Needs ISPs to either provide information or infrastructure changes. - Locality-based approaches do not treat all peers fairly.
Ono - Choffnes and Bustamante (2008)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	ISP clustering. Landmarking based on existing CDN infrastructure (CDN redirection measurements).	+ Needs no ISP cooperation. + Needs no network topology information. - Depends on feeds from major Internet infrastructures and large deployment of its client. - ISP-friendly approaches do not seem to have a great impact on Internet-scale basis.
Liu et al. (2009)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	AS hop count minimization on neighbor selection, on chocking/unchocking mechanisms and on next-chunk picking. Peer selection metric that takes both downloading speed and network topology into account.	+ Optimization of the inter-AS traffic. - Locality based approaches do not treat all peers fairly. + No need for additional infrastructure.
TopBT - Ren et al. (2010)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Applied in multiple places of the BitTorrent protocol (bootstrap, connection establishment/replacement, unchoking).	+ Enhances both traffic and downloading. - Needs off-line processing of BGP dumps.
UTAPS - Li et al. (2008)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Network tomography to construct a picture for the underlying network.	+ Reduced ISP burden. + Better downloading speeds. - Small, laboratory-scale evaluation setup.
Qin et al. (2009)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Cluster peers in a swarm into local, intra- and inter-ISP.	+ Reduced ISP burden. + Better downloading speeds. - Similarly to UTAPS, small scale evaluation setup.
PROP - Qiu et al. (2007)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Neighbor exchange between peers.	+ Cooperation between peers. + Guarantees the connectivity of the network between exchanges. + Can be applied to both fully unstructured and super-peer based architectures. + Secures connectivity of the network.
DDNO - Zeinalipour-Yazti and Kalogeraki (2006)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Domain name topology detection (Split-Hash and dnMatch).	+ Reduces cost of message exchange. + Focuses on both construction and adaptation.
CTAG - Zhao and Lu (2006)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Clustering based on longest matching IP segment.	+ It is independent of the overlay model. + The technique can be considered scalable. - Uses potentially unreliable network latency metric (this can lead to load imbalance etc).

(continued on next page)

Table 1 (continued)

Algorithm / Paper	Topology adaptation	Forwarding optimization	Caching/ Replication	Landmarking	Highlights	Pros/Cons
Landmark Binning - Ratnasamy et al. (2002a)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Landmark binning.	<ul style="list-style-type: none"> - The use of landmark servers renders the technique not fully distributed. - Excessive traffic flow towards the landmark servers is possible. - Fixed points in a network are inherently more exposed to malicious attacks. - Coarse-grained scheme. + Fully distributed. + It is independent of the overlay model.
mOverlay - Zhang et al. (2004)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Dynamic landmarks.	<ul style="list-style-type: none"> + Load balanced. - Coarse-grained scheme.

1039 In what follows, we outline how the surveyed protocols use elements of the above four methodologies. We then offer a summary
 1040 qualitative comparison for all surveyed approaches applicable for unstructured P2P systems. Specifically, Table 1 summarizes, for each
 1041 effort, the methodologies it incorporates, its special highlights as well as a brief presentation of the pros and cons in its implementation.
 1042 Figs. 9–11 pictorially compare the surveyed techniques in terms of efficiency, overhead and scalability; we defined these three criteria in Section 2.3 and we collectively used them as a yardstick to qualitatively evaluate the presented approaches. Last but not least, Fig. 12 creates a timeline of the algorithms and their elements where we capture the temporal and trending dimensions of the surveyed material.

3.2.1. Topology adaptation methodology

Protocols based on topology adaptation modify the topology of the P2P network using various schemes, the two most common of which create either spanning trees using connection graphs or clusters of physically close nodes.

Narada and subsequent algorithms including AOTO, LTM, SBO try to solve the problem by building “a richer connected graph” and by forming minimum spanning trees over this graph that can efficiently route messages among peers. AOTO, LTM and SBO try to overcome this limitation with ingenious schemes like forming minimum spanning trees for the 2-hop away neighbors for each node, separating participant nodes into groups, sometimes with different responsibilities and tasks at hand (e.g., SBO). The advantage of building minimum spanning trees is that they maintain the connectivity on the network in an efficient manner while still preserving the overall search scope. However, their construction and their update costs, especially in dynamic and high-churn environments, may cause large traffic overheads on the underlying network (Chu et al., 2000, 2001, 2002).

The cluster-based approaches, on the other hand, link physically-close nodes to each other. T2MC, for example, uses trace-route logs and DDNO exploits domain names to cluster nodes in proximity at peer join. Further enhancements may include dynamic local restructuring of the overlay graph through neighbor exchange like in PROP or cycle-cut like in DCMP to achieve continuous adaptation throughout a peer's life-cycle. Unfortunately, commonly used methods for proximity detection across Internet do not always return reliable results and therefore, mapping accuracy is not guaranteed. Specifically for traceroute, its overhead is not negligible and routers or firewalls in a network may have already been configured to disable traceroute response from the start. The most problematic aspect of clustering, though, is its nature per se. Limited connectivity among the various local domains can significantly shrink the search scope, negatively affecting the query response time that the P2P user experiences. Among others, DDNO tries to balance the efficiency of the clustering approach with enhanced node connectivity by forcing half of each node's connections to be with other, randomly selected, nodes.

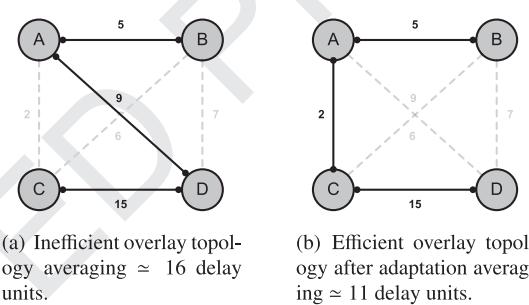


Fig. 5. A simplified example on how overlay topology adaptation may improve matters.

As a methodology, topology adaptation, ultimately aims to reduce the average path traversing cost from one node to another. In doing so, the methodology re-arranges the overlay network so that it becomes a better fit for the underlying IP network as Fig. 5 depicts. Despite the fact that the above is a plausible proposition, it is not always advantageous. For example, a topology adaptive algorithm can exchange a slow edge in the overlay with a faster one thus reducing the average latency of the network communication. The pitfall here is that this new virtual link can traverse a fast AS-to-AS link meaning that even though message round-trip-time is reduced, it has actually additional cost in terms of inter-ISP communication accounting and management.

3.2.2. Forwarding optimization methodology

In unstructured systems, there is no way to know where data are actually located. This is why early works like Gnutella (2000) used blind flooding, a BFS-approach with depth d and a system-maximum TTL value (in terms of hops) for messages. This is inefficient since it generates a large number of messages, a lot of which are duplicates. Approaches based on forwarding-optimization propose intelligent search mechanisms in an attempt to make searching more efficient.

Fig. 6 shows a simple example of how forwarding optimization may work. Fig. 6 (a) shows a protocol that simply floods the entire network in search of object found in the nodes colored in black. Alternatively, a node can decide to forward its messages not to every output link but to a specific or subset of its outgoing links. In Fig. 6(b), the node in the middle, does not flood its neighbors; instead it picks only one of them. The dashed line on the right, shows that, for this routing process, the protocol has rendered two potential forwarding paths as inefficient (e.g., the target nodes show overloading signs).

Many alternative schemes have been proposed for this category, some of which we have surveyed; iterative deepening, modified BFS, local indices (Yang and Garcia-Molina, 2002), k -walker random walks (Lv et al., 2002) or heterogeneity exploitation and token-based flow

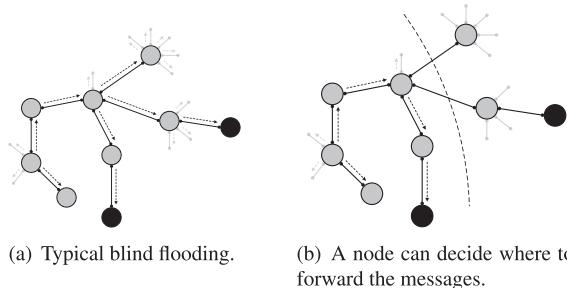


Fig. 6. Forwarding optimization in constraining message flooding.

control mechanisms (Chawathe et al., 2003). More sophisticated alternatives also exist, taking into account application-level requirements or even exploiting machine learning techniques to adjust overlay network routing (Bernstein et al., 2003).

Forwarding-optimization schemes in unstructured P2P systems can be classified as *DFS*- or *BFS*-based. Routing indices (Crespo and Garcia-Molina, 2002) for example is a *DFS*-based technique while all the aforementioned are *BFS*. Also, the schemes in question can be classified as deterministic or probabilistic (probabilistic, random or ranking-based query forwarding). Iterative deepening and local indices (Yang and Garcia-Molina, 2002) are examples of deterministic approaches. Moreover, in the literature algorithms are also taxonomised as blind or informed depending on whether nodes keep some metadata to facilitate the search. For example, *k*-walker random walks or iterative deepening are considered blind searches while local indices informed.

On one hand, forward-optimization approaches have the advantage of enhancing the search responsiveness and reducing the aggregate resource usage of the physical network. On the other, they suffer from drastic reduction of the search scope (in Fig. 6(b) the object on the far right is not reached) thus, limiting the scalability of the whole network. Forward-optimization suggestions address the problem of the mismatch in a limited manner since they do not provide any guarantees that overlay and underlying topologies are aligned with each other. To this end, forward-optimization techniques are commonly applied in conjunction with other methodologies to improve the overall quality of P2P systems.

3.2.3. Caching and replication methodology

Caching is widely used to exploit locality and minimize redundant transfer of data. Caching has with much success been successfully adopted by web- and file-server application environments. Since peers in a P2P system also operate as servers, it is intuitively expected that P2P file-sharing systems can also benefit from caching in improving performance and reducing overall resource usage. However, the design of caches in this context is non-trivial compared to the web-based caching. Due to the fact nodes play the double role of server and client, two important issues have to be considered at design time. First, the lifetime of a query is short, as the nodes join and leave frequently. Second, the result of a single query string is not always the same, as this depends on the source of the query, the TTL value set for the messages, the current interconnection of peers and the high volatility of the environment. Thus, to develop a successful caching system for a P2P architecture, these parameters also have to be carefully considered. P2P caching/replication can be applied at two different levels, namely caching indices or pointers to data (Fig. 7(a)) or caching the data itself (Fig. 7(b)).

Successful implementations have already been developed in some commercial P2P systems, like *KaZaA* or less popular ones including *Gia* and *BNS*. Even though the state of the art in P2P protocols using caching methods helps reduce the burden of network resources, their

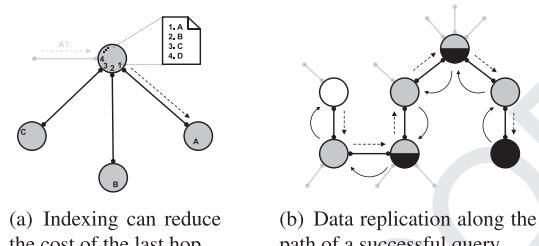


Fig. 7. Index and data replication strategies.

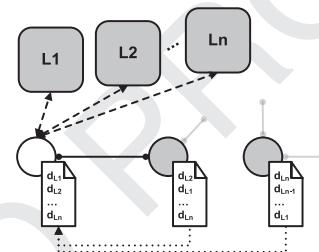


Fig. 8. Landmark binning during node bootstrap.

contribution in addressing the actual mismatch between the overlay and the underlying networks remains limited.

3.2.4. Landmarking

In landmark-based algorithms, nodes use network delay (e.g., RTT) as a distance measurement method to position themselves with respect to "a priori" known servers on the Internet, like *Ono* which uses the CDN-infrastructure for this purpose.

For example, in Fig. 8, the newly arriving peer, measures its distance to an array of landmark points denoted as L_n . Then it creates an ordered list of the landmarks based on its measured distances uses that information to select its neighbors. Specifically, the landmark-servers are used by nodes to estimate their positions based on the intuitive assumption that peers with similar distances to a set of landmarks, are physically close to each other, as well, over the network.

Landmark-based protocols have four important drawbacks: first, the network delay is not a reliable distance estimation method. For example, based on the load on the network, the delay to certain nodes or networks can significantly change over time; this eventually affects the distance measurements and yields wrong measurements that ultimately lead to wrong estimated positions for the nodes or incorrect and non-optimal clusterings of nodes. Second, it can be characterized as a rather coarse-grained approximation, therefore not particularly well suited for identifying the correct positions of nodes within close distance to each other. Third, relying on predefined nodes makes the whole paradigm not fully distributed and the landmark system prone to single point of failure. Lastly, landmark infrastructure requires costly installation and maintenance across the Internet and the different autonomous system domains. As popular P2P file-sharing applications usually have millions of peers connected to at any time, the network costs of maintaining these landmark servers can be quite high. A possible solution to the scalability problem of the static landmark servers is to use ordinary nodes as dynamic landmarks in an approach reminiscent to that of *mOverlay*. Even though this scales better, the accuracy of measurements may affect the overall performance of the system, especially in an environment with high churn.

4. Structured P2P networks

This section offers description and analysis of algorithms proposed to address the topology mismatch problem in structured P2P

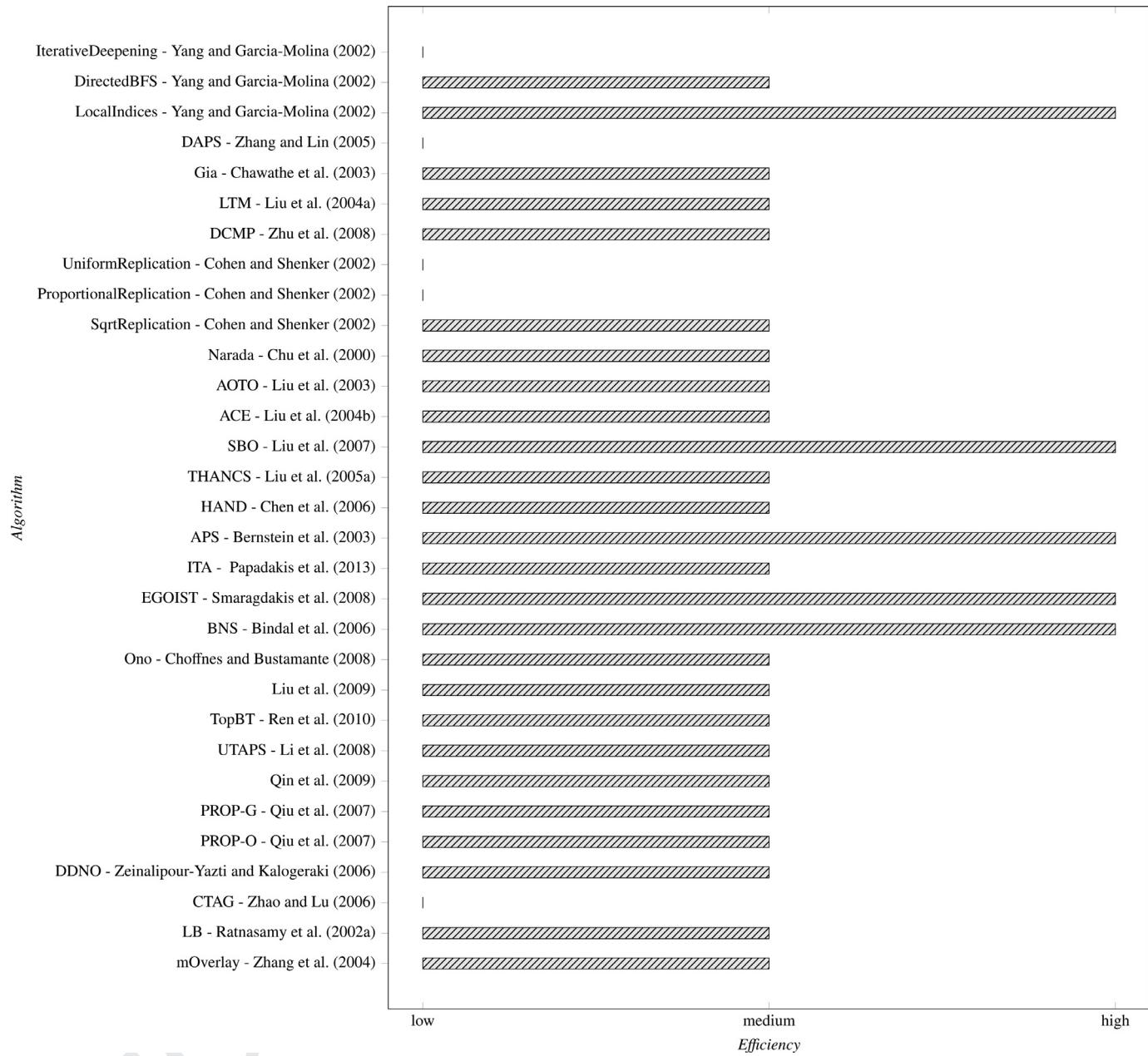


Fig. 9. Efficiency pictorial comparison of unstructured approaches.

networks. The categorization of the approaches is based on how different levels of proximity information is gathered as well as the way peer routing tables are maintained and used in an effort to make overlays as efficient as possible in terms of hops traversed. Our categorization is based on previous work found in (Castro et al. 2002b), (2003), (Ratnasamy et al. 2002b).

4.1. Algorithms for structured architectures

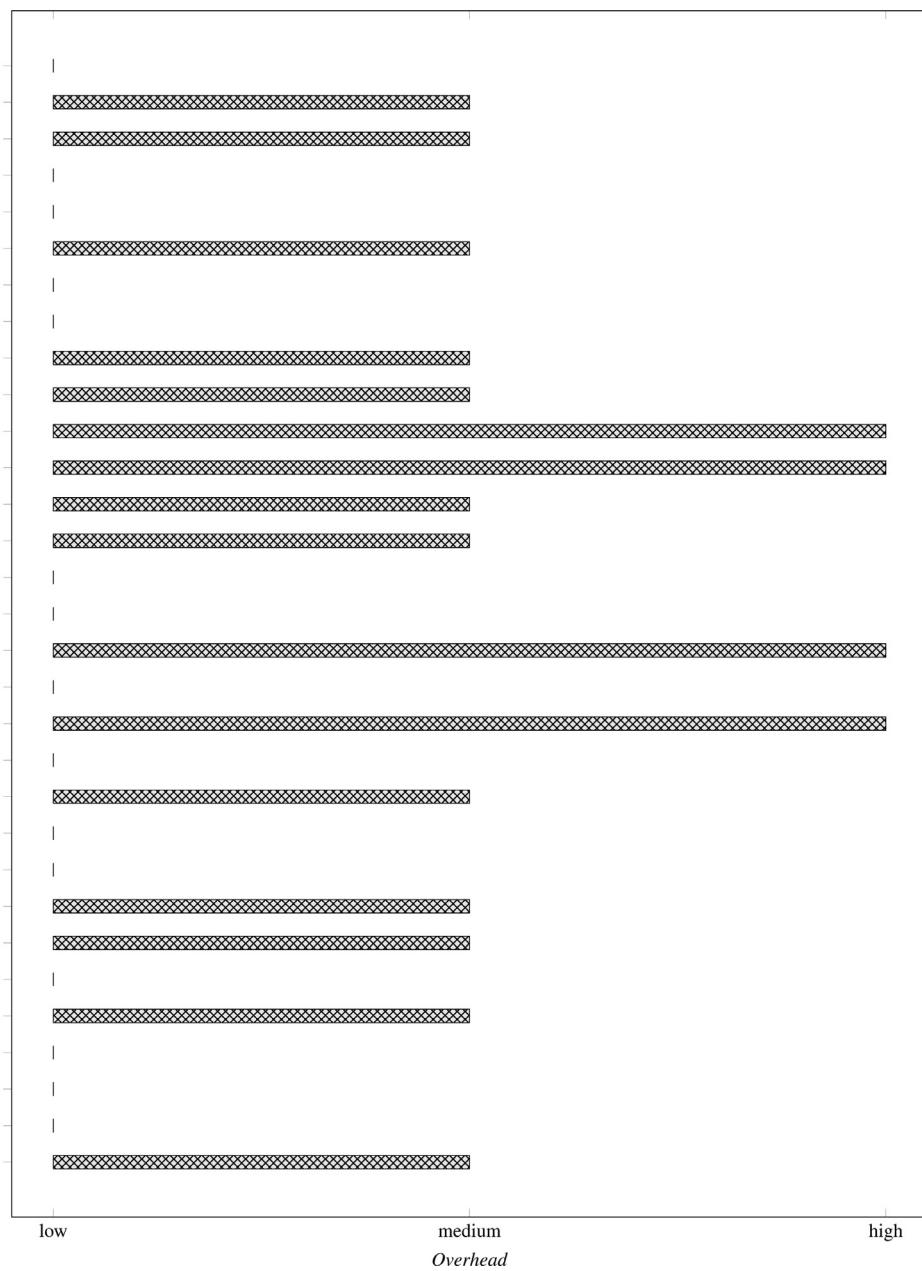
Global Soft-State (Xu et al., 2003) builds a global view to help nodes choose shorter routing paths. To generate the proximity information, it combines the landmark binning method with small scale distance probes to reveal the properties of the underlying network.

This global view of the network state is used by organizing it into "maps" for regions of the overlay, that are strategically placed across the network and made available to any node. The notion of a region is

a well-defined property for structured P2P networks and can be part of the Cartesian space in overlays like eCAN (Xu and Zhang, 2002) or a set of nodes sharing a particular prefix in overlays such as Pastry (Rowstron and Druschel, 2001). Moreover, a node may appear in a maximum of $\log(N)$ such maps.

Direct access to this information, also coined soft states, helps nodes find the best way to route their messages, while a publish/subscribe mechanism allows them to get notified about dynamic network changes.

Although this approach does reduce the routing latency to far nodes, it can become expensive as it is possible to end up with a very large number of regions each of which has to maintain a map. The fact that a peer may appear in multiple such maps incurs additional traffic for probing and notifications and so renders the approach unscalable (Ren et al., 2004). Moreover, to refine measurements additional inner-bin measurements are needed (Winter et al., 2004). Also, improvements within a region can be minimal as there exists limited

**Fig. 10.** Overhead pictorial comparison of unstructured approaches (lower is better).

knowledge about neighboring zones. In terms of our three stated criteria, the approach fares as follows:

Global soft-state	Efficiency	Overhead	Scalability
	Medium	High	Low

Mithos (Waldvogel and Rinaldi, 2003) ensures that neighboring nodes in the overlay are close also in terms of the underlying layer by integrating proximity layout and proximity routing overlay optimizations.

During bootstrap, an incoming peer J obtains from a bootstrap node a set of candidate peers with which to establish links. Subsequently, the node works iteratively to identify peers that minimize a given metric (i.e., network delay). The process stops when no additional peers complying with the above metric can be found. To tackle the local minima problem that such a gradient descent method suffers from, Mithos probes nodes within a 2-hop radius from the cur-

rent node that helps minimize the distance before concluding the process.

J is then assigned a synthetic ID that represents the state of minimum distance from its nearby 2-d neighborhood discovered during the aforementioned phase. The benefit of adopting synthetic coordinates is that distance computations among nodes can be done in the ID space without requiring physical measurements. Mithos then proceeds to designate J 's closest neighbors as previous steps do not reveal with certainty which nodes are its closest neighbors. Using Mithos quadrant-based mechanism, J divides the 2-d space of neighbors in four regions and subsequently, establishes a link to the closest neighbor in each quadrant of the plane. It does so, by starting from a known node in the neighborhood and scanning towards all quadrant borders (Karp and Kung, 2000; Rinaldi, 2002).

A noted Mithos limitation is that the protocol cannot effectively handle dynamic arrivals and departures from the network (Ren et al., 2004). Costa et al. (2004) points out extensive probing for Mithos to

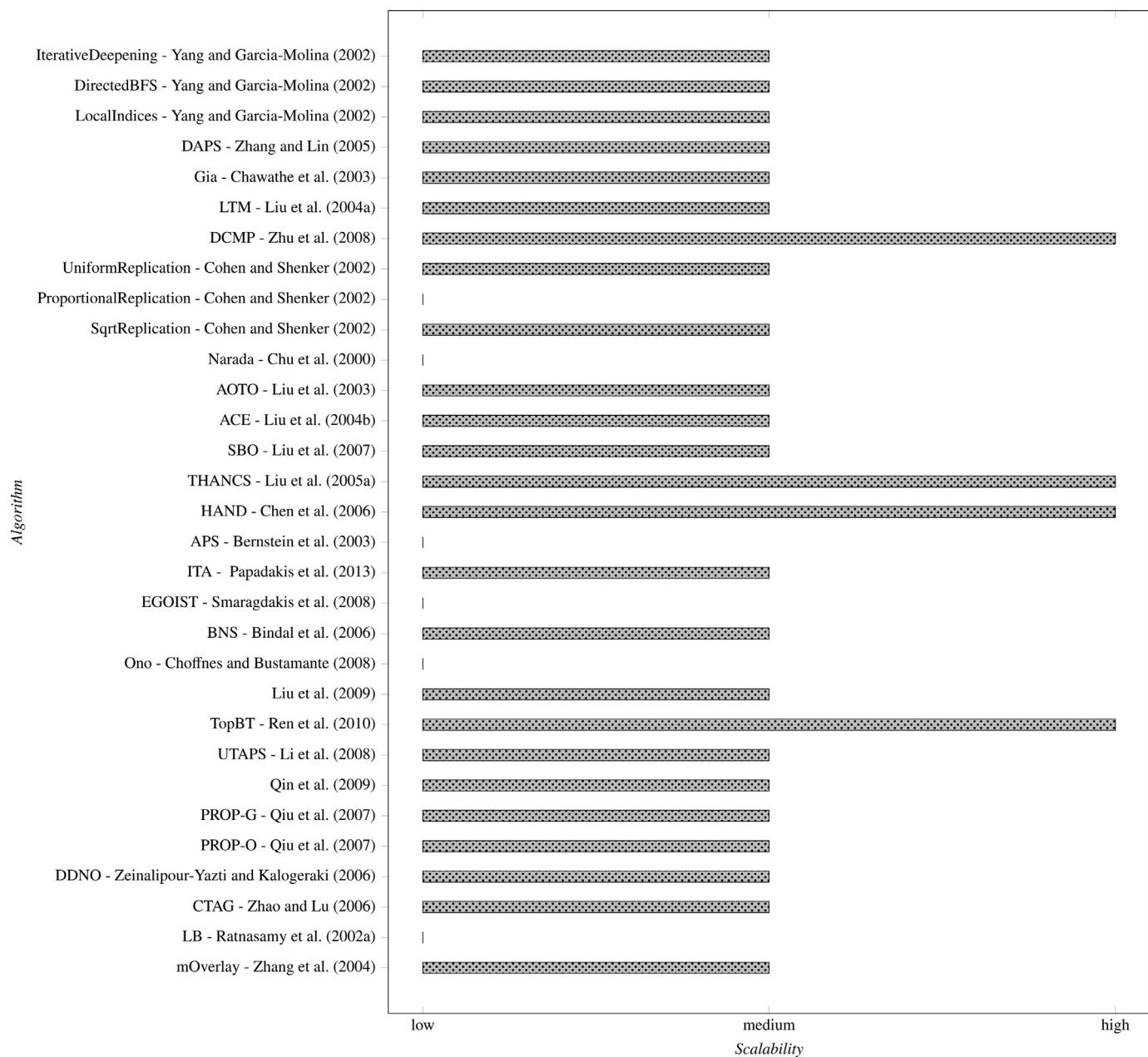


Fig. 11. Scalability pictorial comparison of unstructured approaches.

1273 determine distances and less accurate proximity results compared
 1274 to other virtual coordinate systems. The last issue has been also at-
 1275 tributed by Wong et al. (2005) to the iterative neighbor selection pro-
 1276 cess that is reportedly prone to premature termination at local min-
 1277 ima. Cox et al. (2004) argues that elements of the protocol's synthetic
 1278 ID generation require a centralized implementation affecting *Mithos'*
 1279 scalability. In regard to the three criteria, *Mithos* stands as follows:

	Efficiency	Overhead	Scalability
<i>Mithos</i>	Medium	Medium	Low

1280 **LAPTOP** (Wu et al., 2007) organizes the overlay into a tree-based
 1281 hierarchy so that hops required during message routing are reduced
 1282 and maintenance overhead is minimized. A caching scheme is also
 1283 used to further reduce routing table update costs. It is theoretically
 1284 shown that *LAPTOP* routing path length is bounded by $O(\log_d(N))$
 1285 and node joining and leaving in the overlay network is bounded by
 1286 $O(d * \log_d(N))$ hops in a balanced overlay tree with N representing the

1287 number of nodes and d the maximum degree of each node (i.e., num-
 1288 ber of links). *LAPTOP* adopts the geographical layout approach and
 1289 constructs its layout in a self-organizing and efficient fashion. The
 1290 physical distances are estimated with round-trip times to a few ex-
 1291 isting nodes in the overlay network. Each node is assigned a dot-
 1292 formatted address (e.g., 1.3.4) with each octet ranging from 1 to d .
 1293 The assignment process is done by appending a unique octet to the
 1294 address of the parent of each node, while the root node is assigned
 1295 address 1.

1296 *LAPTOP*'s routing scheme is similar to the longest-prefix IP match-
 1297 ing scheme. At each forwarding hop, a message travels up the tree
 1298 until the first common ancestor of both source and destination nodes
 1299 is reached and then, the message starts descending towards its target.
 1300 During tree traversals, special entries in the routing tables are main-
 1301 tained to increase routing efficiency and achieve finer-load balance.
 1302 These special entries constitute the *routing cache* which enables a
 1303 node to forward a message to a better longest-prefix match than that

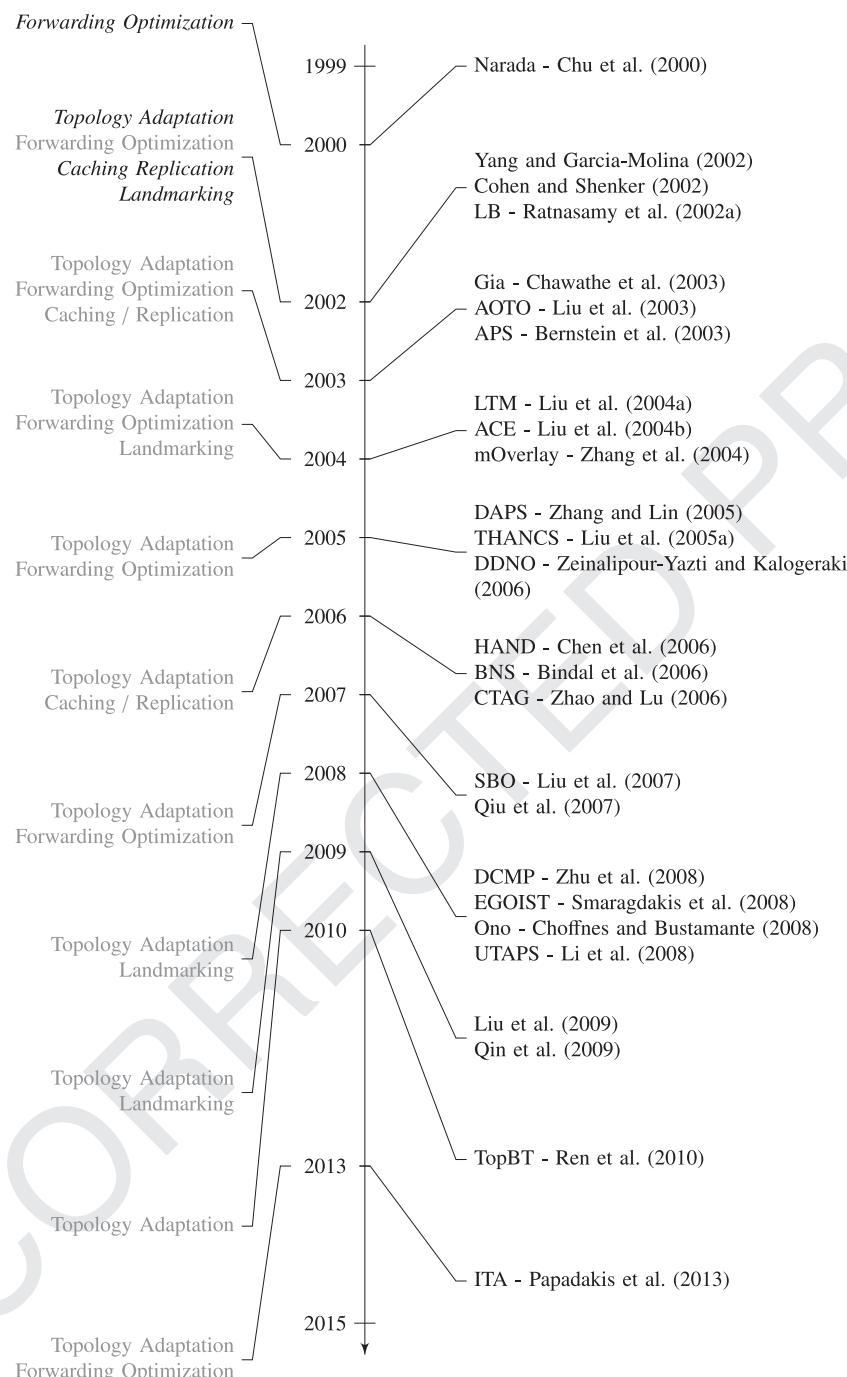


Fig. 12. Trends on characteristics of unstructured approaches (with emphasis denoting first appearance).

of its direct ancestor; this yields a quicker and more cost-effective step through the overlay towards the destination. To improve scalability, the number of child nodes and the size of the routing cache are limited. To contain maintenance costs, *LAPTOP* employs a simple heartbeat-based regime in which a parent is responsible for the monitor of its own children. *LAPTOP*'s behavior for the three criteria stands as follows:

	Efficiency	Overhead	Scalability
<i>LAPTOP</i>	Medium	Low	High

To optimize the underlying network use, [Kaune et al. \(2008\)](#) introduces proximity controls over the *vanilla Kademlia* ([Maymounkov and Mazières, 2002](#)) protocol's random *ID* assignment. In its quest to

improve routing performance, the overlay uses a cost "yardstick" (i.e., metric) provided by the underlying network. Such cost information can be acquired by: (i) measuring previous lookups, (ii) exchanging or jointly-computing measurements with others, or (iii) using a local database to look-up information.

[Kaune et al. \(2008\)](#) exploits two routing techniques: *proximity routing selection (PRS)* and *proximity neighbor selection (PNS)*. In *PRS*, the goal is to choose the best next hop while routing a message. *Kademlia* always selects the closest-node with respect to the XOR metric ([Maymounkov and Mazières, 2002](#)). With *PRS* in place however, *Kademlia* may also choose the node that exhibits the smallest routing cost; in this way, it balances a trade-off between the logical overlay and actual underlying distances. In *PNS*, on the other hand,

1327 the goal is to keep peers with the least routing cost in the routing
 1328 table.

1329 *Kademlia* proximity injection improves locality of connections
 1330 among peers. To enhance traffic locality, the use of *MaxMind GeoIP*
 1331 database⁴ is suggested to form clusters of nearby peers. To this end,
 1332 Vivaldi coordinates (Cox et al., 2004) could be also used as an alter-
 1333 native for morphing locality. Kaune et al. (2008) however suggests
 1334 that the clustering approach when combined with PRS/PNS helps con-
 1335 tain 40% of messages within an ISP and allows only 6% to go through
 1336 the backbone. The behavior of Kaune et al. (2008) with respect to the
 1337 stated criteria is deemed as follows:

	Efficiency	Overhead	Scalability
Kaune et al. (2008)	Medium	Medium	Medium

1338 **CHOP6** (Morimoto and Teraoka, 2007) is a Chord-variant DHT that
 1339 exploits IPv6 address format as well as available RTT information to
 1340 enable proximity estimation among peers during message routing.
 1341 Similar to *Chord* (Stoica et al., 2001), CHOP6 uses a finger table with
 1342 each entry now holding $k > 1$ candidate nodes. The next hop for a
 1343 message is selected to be the one closest to the destination among
 1344 those k candidate peers. This proximity is determined either by ex-
 1345 ploiting RTT measurements or, if there is no such information yet (i.e.,
 1346 a new peer appears), by estimating the distance in the overlay's ID
 1347 space.

1348 When joining the network, each node receives a 64-bit *ID*. It com-
 1349 prises of a random part and the IPv6 prefix which is obtained by
 1350 hashing the node's IP address and port number; this combination
 1351 is deployed in an effort to preserve the uniformity of node assign-
 1352 ment. The prefix in question is based on IPv6's global routing prefix.
 1353 As IANA⁵ assigns address blocks to regions (RIRs) which are further
 1354 distributed at national level (NIRs), the first 32-bits of the address can
 1355 estimate the region or the country in which the node is connected.
 1356 Thus, CHOP6's IPv6 prefix ID is used for proximity estimation if no
 1357 RTT information is available during routing. With respect to the three
 1358 qualitative criteria, CHOP6 stands as follows:

	Efficiency	Overhead	Scalability
CHOP6	Low	Low	Medium

1359 **T2MC**'s (Shi et al., 2008) objective is to diminish the redundant
 1360 multiple messages that may unnecessarily cross autonomous system
 1361 (AS) boundaries. Inter-AS links are the costliest to ISPs and hence,
 1362 minimizing topology mismatches greatly affects pertinent opera-
 1363 tional expenditure. To achieve this, each node exploits the stable
 1364 structure of the Internet routers to help solve a clustering problem.
 1365 The algorithm uses *traceroute*-logs to detect inter-AS boundaries and
 1366 clusters peers to either nearby or far-away groups by applying a cus-
 1367 tomized k -mean classification ($k = 2$). When joining the network, the
 1368 peer randomly picks an IP and uses *traceroute* and selects the routers
 1369 with minimum and maximum latencies in the above *traceroute* path
 1370 as initial centroids for the "nearby" and "far-away" sets respectively.
 1371 Following an iterative procedure, the node comes up with the two
 1372 sets demonstrating minimum intra-set latency variance. At this stage,
 1373 routers in the "far-away" set are not considered critical. In contrast,
 1374 the router from the "near" set having the maximum number of hops
 1375 from the just-joining peer is key as it denotes an edge-gateway. Peers
 1376 that share the same edge-gateway or other members in the "nearby"
 1377 router cluster are considered intra-AS neighbors.

1378 Even though *traceroute* provides detailed information about the
 1379 network structure, its extensive use may create overhead on the over-
 1380 all network structure. Administrators thus frequently disable *tracer-
 1381 route* support on their routers and this undoubtedly affects T2MC's
 1382 operation. Although overall experimental results have been encour-

1383 aging, T2MC is considered rather coarse-grained as it only classifies
 1384 neighbors into two groups (Qin et al., 2009). T2MC fares as follows:

	Efficiency	Overhead	Scalability
T2MC	High	High	Low

1385 **PChord** (Hong et al., 2005) is based on the *Chord DHT* and in-
 1386 corporates proximity in its routing mechanism. To achieve this, the
 1387 standard finger table is augmented with a *proximity list* that contains
 1388 nodes in the network found nearby in terms of RTT. In this respect,
 1389 the next hop during routing can be decided based on both the direc-
 1390 tion towards the search key and minimizing the cost of that hop.

1391 When a peer joins the network, its proximity list is empty. This list
 1392 is maintained in a heuristic fashion during the node's lifetime. When
 1393 a node needs to route a message, it chooses as the next hop the node
 1394 closest to the target key in *both* its finger table and proximity list.
 1395 An initial choice for the next hop made out of the finger list can be
 1396 substituted by a shorter one from the proximity list if it is both closer
 1397 to the key (ID-wise) and closer in physical distance. This yields an
 1398 equal or even more efficient routing than traditional Chord.

1399 Unfortunately the overlay maintenance cost is somewhat higher
 1400 than *Chord* as additional heart-beat messages are required for con-
 1401 nection verification for the entries of the proximity list. Fortunately,
 1402 such entries represent nodes already close-by so the additional cost
 1403 is kept to a minimum. However, PChord has been reported to exhibit
 1404 slow convergence speed and inefficiency in high-churn environments
 1405 (Dao and Kim, 2006). PChord qualitatively behaves as follows:

	Efficiency	Overhead	Scalability
PChord	Medium	Medium	Medium

1406 **AChord** (Dao and Kim, 2006) exploits IPv6's *anycast* mechanism
 1407 to relieve the protocol from complex joining procedures, harness net-
 1408 work proximity and achieve high routing efficiency. When sending a
 1409 message to an IPv6 anycast group, the originator accomplishes this
 1410 delivery by interacting with the physically closest node in that group
 1411 (Metz, 2002). Thus, AChord organizes all nodes participating in the
 1412 overlay network into a single anycast group. An arriving peer is auto-
 1413 matically forwarded to the physically nearest node for bootstrapping;
 1414 this curtails the need to explicitly maintain participants and the effort
 1415 of locating the physically nearest.

1416 The *ID* of an incoming node is computed based on both the iden-
 1417 tifier of the bootstrap node and the bootstrap's predecessor; this is
 1418 done in a way that the *ID* of the incoming peer is positioned between
 1419 the two identifiers mentioned above. After joining, a finger table is
 1420 created as in *Chord*. Moreover, an *adjacency table* maintains infor-
 1421 mation about the k closest-known peers. The routing decision takes
 1422 place in a way reminiscent to *Chord* and uses both adjacency and fin-
 1423 ger tables. AChord behaves as follows:

	Efficiency	Overhead	Scalability
AChord	Medium	Medium	Medium

1424 **Chord6** (Xiong et al., 2005) is another *Chord*-variant that uses
 1425 IPv6's hierarchical features to create a substrate that reduces inter-
 1426 domain traffic among service providers. The key difference between
 1427 *Chord6* and *Chord* lies on the identifier definition. The adopted ap-
 1428 proach is deemed as easily portable to other *DHTs* including *CAN*,
 1429 *Pastry* and *Tapestry*. *Chord6*'s identifier contains two parts: the higher
 1430 bits are obtained by hashing the node's IPv6 address prefix (of spe-
 1431 cific length), while the remaining lower bits of the identifier is the
 1432 hash-value of the remaining of the IPv6 address. As a result, nodes
 1433 in a domain are mapped onto a continuous key space on the overlay
 1434 network; the latter avoids unnecessary message-forwarding across
 1435 different service providers and helps minimize overall routing cost.

1436 *Chord6* reduces the average length of hops traversed to be loga-
 1437 rithmically proportional to the number of available domains instead
 1438 of the number of participating peers. Due however to the large num-
 1439 ber of Internet domains and the possibility of occasionally having

⁴ <http://www.maxmind.com>.

⁵ <http://www.iana.org/>.

only a small number of peers within each domain, **Dao and Kim (2006)** challenges the approach's scalability and efficiency. Overall, the approach fares as follows:

	Efficiency	Overhead	Scalability
Chord6	Low	Low	Low

A proximity neighbor selection scheme functioning atop the *Chord DHT* is introduced by **Duan et al. (2006)** (**DHT-PNS**). Its key objective was to group physically-close nodes as neighbors in the *DHT* table while exploiting proximity information. To detect proximity, virtual network coordinates for peers are used with the help of the Vivaldi protocol (**Cox et al., 2004**). The virtual coordinates are then used to help map nodes to the identifier-space of the *DHT*. The space in question is partitioned using a *concentric circle clustering scheme* where successive cycles of radiiuses $\rho, 2\rho, 3\rho$ and so on, are constructed. The formed annuluses are then divided into $2\chi - 1$ sectors, where χ denotes the sequence number of the annulus starting from $\chi = 1$ for the center cycle. In this way, **Duan et al. (2006)** prove that for each sector this characteristic favors a more load-balanced clustering operation assuming uniform node distribution. Every sector in this $2d$ space is mapped to a unique *region* in the *DHT* space forming a multi-layer node identifier space. Consequently, nodes from the same sector are mapped to the same region as well, preserving their proximity relationship.

Individual elements are uniquely mapped to the identifier-space, allowing logarithmic lookup operations with high probability on *Chord*. To this end, a node can obtain its *DHT* key and its region key, in a fully distributed manner, just by applying a consistent hash function. The node can then communicate with the region's master node called *Cluster Node (CN)* which is responsible for clustering the nodes belonging in the same sector or region with that of the node at hand. The node registers to its corresponding region and publishes its information, through the special peer *CN*; the peer can also ask *CN* for other nodes that have previously joined the region to enhance the peer's neighbor set for future routing table optimization. If no other nodes are found within the region, the search is expanded to neighboring regions. Even though scaling appeared promising, only limited simulations were carried out with configurations involving up to 2048 nodes. Moreover the approach assumes uniform node distribution which is not always the case, especially in a synthetic coordinate system. *DHT-PNS* fares as follows in reference to the stated three criteria:

	Efficiency	Overhead	Scalability
DHT-PNS	Medium	Medium	Medium

Sun and Zhang (2008) propose **Quasi-Chord**, a *Chord*-based network that constructs its overlay network while attempting to match the underlying network topology. The overlay creation follows 3 stages: first, each host acquires its coordinates in a geometric space using the *Global Network Position (GNP)* protocol (**Ng and Zhang, 2001**) and a set of landmarks. Second, using the *Cantor* space-filling curve, the 2-dimensional space is converted to 1-dimensional space. The latter is employed in the third and final stage to build the *Quasi-Chord* circle according to the host's *Cantor* value. Physical proximity is thus denoted by *Cantor* value proximity. The *Chord* cycle is constructed by sorting nodes in ascending order. *Quasi-Chord*'s qualitative behavior is as follows:

	Efficiency	Overhead	Scalability
Quasi-chord	Medium	High	Medium

Proximity neighbor selection algorithms often use probing to detect proximity. However, such methods suffer by imprecisions often due to network overload. **IP-based clustering** or **IPBC** (**Karwaczynski and Mocnik, 2007**) uses IP address prefixes (16 bits for IPv4) to detect proximity. Such proximity information can be easily stored in the *DHT* itself.

Each peer in the overlay has a unique *ID* and an IP address to characterize it. To publish itself, the peer hashes a fixed-length prefix of its IP to obtain a key and stores the resulting *ID* and IP in the *DHT*. Newly joining nodes with the same IP prefix can, thus, query the *DHT* and readily identify all neighbors with the same prefix. Nodes periodically update their entries in the *DHT* by removing those who leave or have timed-out.

The overhead of the algorithm is very low as proximity detection requires only a comparison of prefixes. On the other hand, the efficiency of the proximity is caught into a vicious cycle with the storage space needed for *DHT* object publishing. Moreover, its static nature assumes high correlation of common IP prefix length of participating peers and latency (with no other measurements), which might not always be the case, especially in cases of high-churn environments. **Eom et al. (2011)**. The suggested *IPBC* approach fares as follows:

	Efficiency	Overhead	Scalability
IPBC	Low	Low	Medium

Dynamo (**Winter et al., 2004**) creates an overlay that not only considers physical proximity amongst peers but also takes into account mobility attributes of nodes. To adapt to mobile ad-hoc networks, *Dynamo* strives to maintain an evenly-distributed overlay *ID*-space so that hot-spots are avoided. Although based on *Pastry* and its well understood built-in locality heuristics (**Castro et al., 2002a**), *Dynamo*'s own mechanisms are *DHT*-independent and so, they can be applied to other overlays. *Dynamo*'s approach may be viewed as bottom-up.

An incoming node computes physical information about its neighborhood and uses the outcome to unilaterally assign itself an *ID*. The approach suggests that for every new overlay routing-hop the physical distance between nodes is likely to increase resulting in a dominating last routing step (**Rowstron and Druschel, 2001; Castro et al., 2002a**). To identify this dominating step, two techniques are used: **Random LandMarking (RLM)** and **Closest Neighbor Prefix Assignment (CNPA)**. In *RLM*, a set of peers are chosen to be responsible for a set of landmark keys. These keys have to be chosen in a way that they equally divide the overlay space (i.e., in a 16-base an appropriate split would be 000..0, 100..0, 200..00, ..., F00..00).

An incoming node then computes distances from those landmarks, sorts them, and assigns itself an *ID* on this landmark ordering as follows: the peer sets the prefix of its *ID* to be equal to that of its closest landmark. The length of this prefix is $l_{prefix} = \lfloor \log_b k \rfloor$, where b is the *ID*-base and k the number of landmarks. The remaining of bits of the peer-*ID* are assigned either randomly or computed with additional proximity optimizations. In this way, two desirable features are attained: (i) landmark assignment is dynamic, and (ii) the set of a node leaf may reference peers that are physically close by. The latter in particular helps reduce average routing cost(s) as the leaf-set of a node is used to efficiently determine the last routing step.

CNPA is proposed for network setups for which a lightweight bootstrap phase is required. It delegates the task of identifying the physically close-node in a *Pastry*-based network; this node is used by the incoming peer to get bootstrapped. The new-comer takes the bootstrapping node's *ID* prefix and fills the remaining bits of its key using the *RLM* technique. *RLM* is evaluated and exhibits good proximity in load-balanced environments. Moreover, it is adaptable to high-churn settings. Unfortunately, *RLM* bootstrap overhead is substantial and can effect the set of landmarks (**Winter et al., 2004**). *CNPA* on the other hand, incurs less overhead but this comes at the expense of being more coarse-grained as it is based on prefixes to determine proximity during bootstrap. *Dynamo*'s elements fare as follows regarding our three criteria:

	Efficiency	Overhead	Scalability
RLM	Medium	Medium	High
CNPA	Medium	Low	Medium

1555 **3D Overlay (3DO)** (Abid et al., 2014) seeks to address the network
 1556 mismatch problem for peers operating in mobile networks. As its
 1557 name suggests, the network logical space is a 3-d space and the used
 1558 “transient” (continually changing) logical IDs (*LIDs*) consist of 3 axes,
 1559 x , y and z in the form of a tuple; each such axis value is represented
 1560 by an M -bit identifier. A newly-arriving peer broadcasts a join request
 1561 using a TTL-based expanding ring search helping locate peers that are
 1562 physically adjacent. Responding peers send in their *LIDs* along with a
 1563 list of their neighbors and their corresponding distances from those
 1564 neighbors. Once a joining peer receives the above responses, it stores
 1565 requisite information in its own routing table and so, it generates an
 1566 undirected connected graph of itself with both its 1- and 2-hop neigh-
 1567 bors. Starting off this graph and with the help of hop distances ob-
 1568 tained from the underlying MANET, the incoming peer builds a min-
 1569 imum spanning tree (MST) with itself as the root node; in this MST
 1570 the nodes adjacent to the root become the direct neighbors of the
 1571 peer. Moreover, the entries of the initial routing table that do not cor-
 1572 respond to above direct neighbors are eliminated.

1573 The above procedure ascertains that close peers on the overlay are
 1574 also adjacent at the physical layer. At this stage, a peer can generate
 1575 its *LID* that automatically places it in the 3-d network space. Updates
 1576 in the peer’s connectivity information reflecting continual changes in
 1577 the mobile network, trigger updates on the composition of *LID* itself.
 1578

1579 **3DO** stores objects in the same 3-d space. For each file f , a key k
 1580 of the form of an x , y and z tuple is hashed out of the file’s unique
 1581 *ID* (e.g., filename). The node whose *LID* is closer to k stores the file
 1582 itself. To retrieve the file, the requesting node calculates f ’s k and is-
 1583 sues a query message. The message is associated with a TTL value.
 1584 Each peer forwards that query to the node, in its routing table, that is
 1585 closest to k until either the peer that stores the file is located or the
 1586 query’s lifetime expires.

1587 **3DO** focuses on the creation of a 3-d space whose stated objective
 1588 is to yield improved load balancing and exploit fine-grained informa-
 1589 tion to ensure that each hop on the overlay during object location, is
 1590 also closer physically to the sought object. However, the use of TTL in
 1591 query messages renders the network not-a-pure DHT as it abolishes
 1592 the zero false-negative queries that traditional DHTs maintain. Abid
 1593 et al. (2014) also concedes that the approach does not rapidly adapt
 1594 to network changes and so its evaluation was solely based on low
 1595 mobility scenarios. For the above reasons, **3DO** exhibits the following
 behavior with respect to our three stated criteria:

	Efficiency	Overhead	Scalability
3DO	Low	Low	Medium

1596 **SAT-Match** (Ren et al., 2004) defines an iterative process that
 1597 adaptively changes the overlay network to ease the topology mis-
 1598 match problem. It consists of a *probing phase* in which a peer queries
 1599 nearby nodes for distances and a *jumping phase* in which the peer
 1600 decides to change its neighborhood to reduce average neighbor dis-
 1601 tances. This process is performed iteratively until no additional opti-
 1602 mization is required; yet, the methods achieves optimization within
 1603 a sufficiently large scope. The notion of *stretch* is used as the means
 1604 for quantifying the *topology match degree* of the constructed overlay;
 1605 stretch is defined as: $S = \bar{L}_l / \bar{L}_p$ where \bar{L}_l is the average logical link lat-
 1606 ency and \bar{L}_p is the average physical link latency. When a peer joins
 1607 the *DHT*, it starts its *probing phase* by contacting the neighborhood
 1608 with messages having small TTL = k (Jiang et al., 2008). Recipients
 1609 return information to the source and forward the message if not ex-
 1610 pired. After discovering its TTL- k neighborhood, the source measures
 1611 RTT-times, sorts them and picks the two nodes with the smallest RTTs.
 1612 During the *jumping phase*, the source node calculates the stretch
 1613 change to all its direct neighbors and to the direct neighbors of the
 1614 first peer that was selected in the probing phase, as if the jump has
 1615 been performed. Using heuristics, the node determines if the stretch
 1616 reduction is sufficient to perform the actual jump; otherwise, it se-
 1617 lects the second peer from the probing phase and repeats the compu-

tation. If again, the threshold is not met, then no jump is ultimately
 1618 carried out.

1619 Experimentation shows that *SAT-Match* attains a 27% stretch
 1620 reduction over a landmark-enhanced CAN-alternative. This result
 1621 shows that *SAT-Match* may perform better than landmark binning
 1622 in terms of the matching degree of the logical to the physical net-
 1623 works. *SAT-Match* behaves as follows with respect to the stated three
 1624 criteria:

	Efficiency	Overhead	Scalability
SAT-Match	High	Medium	Medium

4.2. Discussion on the algorithms for structured architectures

In this section, we have discussed research efforts trying to address the issue of inefficient network resource utilization manifested by network-topology mismatches in *structured P2P*-systems. Based on the salient feature of outlined techniques (Castro et al., 2002b; 2003; Ratnasamy et al., 2002b), we classify furnished solutions based on the following aspects:

- (i) geographic-layout,
- (ii) proximity routing, and
- (iii) proximity neighbor selection,

Trends for structured systems depicted in Fig. 19 show the same lack of paradigm shift we identified for the unstructured approaches (check Fig. 12). Leveraging proximity information for selecting the next hop or to construct and maintain routing tables as a whole, are the two predominant elements used throughout the years in the structured systems literature. Similarly to unstructured approaches, most of the methods reviewed in this section entail features from multiple of the above categories. For example, Kaune et al. (2008) uses both proximity routing and proximity neighbor selection to achieve topology adaptation, while LAPTOP combines geographic-layout with replication strategies to refine its results. What looks different compared with the trends in unstructured systems is the fact that, here, researchers seem to juggle with more methods at the same time. This, and the fact that these suggested solutions are also heuristic in nature, opens them up to trade-offs that are more difficult to manage and are highly affected by both the application domain and the implementation itself.

In what follows, we place the surveyed *structured P2P*-methods in the three above categories formulated based on utilized proximity and geographic aspects. Table 2 outlines key properties for all surveyed approaches that helps differentiate and better position competing techniques. Clearly, we do not further elaborate topics that we have included and discussed at length in Section 3; these aspects entail topology adaptation, landmarking, and caching/replication. Figs. 16–18 help with the pictorial comparison of the algorithms’ behavior with respect to the three criteria we defined in Section 2.3. Finally, we embed algorithms and characteristics on the temporal axis in Fig. 19 to help depict emerging trends during all these years of research.

4.2.1. Geographic-layout

This category consists of methods that try to map the overlay id space on to the physical one, as depicted in Fig. 13. Routing tables of involved peers, are adjusted and maintained by exploiting proximity information that describes the overall (geographic) positioning of the participants. This comes in contrast to proximity routing and proximity neighbor selection (discussed later on) where decisions are taken based on and affecting a peer’s nearby environment.

A number of the reviewed algorithms including *Global Soft-State*, *Mithos* and *LAPTOP* resort to clustering nearby nodes to reduce the network’s average delay. Landmark servers and RTT measurements are two popular methods, which can also be used in conjunction, to

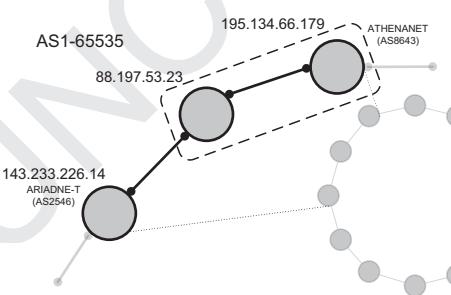
Table 2

Summary table for structured algorithms.

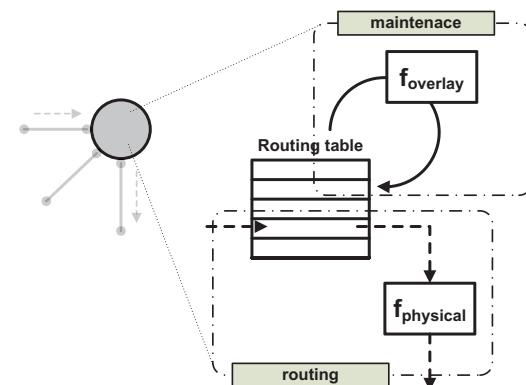
Algorithm / Paper	Topology adaptation	Landmarking	Proximity routing	Proximity neighbor selection	Geographic layout	Caching/ Replication	Highlights	Pros/Cons
Global Soft-State - Xu et al. (2003)	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Hybrid landmark binning and probing scheme for proximity detection. Strategic injection of proximity information across the network. Subscription-notification system to dynamically adapt to network changes.	+ Greatly reduces routing latency to far away nodes. – Potential maintenance costs are high. – Little knowledge about neighboring regions. – Unable to identify nodes that are close to routers/gateways. – Static nature due to the use of landmarks. + Distance measurement is done on the overlay level. – Does not effectively manage dynamic peer arrivals and departures. – Centralized implementation of its spring relaxation technique implementation
Mithos - Waldvogel and Rinaldi (2003)	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Directed incremental probing. Synthetic coordinates.	+ Reduces hops during message routing. + Minimal maintenance overhead. – Heartbeat approach incurs overhead even when not needed. + Proximity routing works in Kademlia to improve connection locality. – Using virtual coordinates renders the approach rigid in high churn environments.
LAPTOP - Wu et al. (2007)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Tree-based hierarchy.	+ Relatively straightforward integration of IPv6 global routing prefix into ID space. – Coarse-grained.
Kaune et al. (2008)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Replacing XOR metric with a function that minimizes the underlying cost. Clustering (MaxMind geolocation technology). Ipv6 format exploitation.	+ Prioritize interaction of peers and edge gateways. – Coarse-grained.
CHOP6 - Morimoto and Teraoka (2007)	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Additional RTT information.	- IPv6 is still not widely applied and supported. + Can constrain costly jumps in and out of network partitions.
T2MC - Shi et al. (2008)	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Clustering using trace-route logs.	+ Maintenance cost kept to a minimum (node join/leave heartbeat for lists). – Rigid in high-churn environments. + Delegates proximity calculation during bootstrap to the anycast mechanism. – Trace-route adds overhead and is sometimes disabled by ISPs.
PChord - Hong et al. (2005)	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Maintenance of a proximity list.	+ Easy to implement (just the additional neighborhood table). – No adaptation after node bootstrap. + Reduces inter-domain traffic between ISPs. + Easily portable to other DHTs. – Cannot distinguish between nearby domains. – Not adaptable after id assignment
AChord - Dao and Kim (2006)	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Exploits IPv6's anycast mechanism.	+ Good convergence speed and optimizing performance. – It assumes uniform node distribution which is not always the case. – Not fully distributed (GNP is landmark based).
Chord6 - Xiong et al. (2005)	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Exploits IPv6's hierarchical features.	(continued on next page)
DHT-PNS - Duan et al. (2006)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Grouping through synthetic coordinates. Partitioning using a concentric cycle clustering scheme. Geometric space coordinates (GNP)	

Table 2 (continued)

Algorithm / Paper	Topology adaptation	Landmarking	Proximity routing	Proximity neighbor selection	Geographic layout	Caching/ Replication	Highlights	Pros/Cons
Quasi-Chord - Sun and Zhang (2008)	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Transformation of the coordinate space into 1-d Cantor space for easier mapping to the Chord hierarchy Two finger tables (clockwise, anti-clockwise)	- Makes an indirect assumption of a maximum number of allowable hosts. - Doubles the required routing information which needs to be created and maintained.
IPBC - Karwaczynski and Mocnik (2007)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	IP address prefixes (16 bit for IPv4) to detect proximity.	+ Prefix is stored in the DHT so the proximity identification becomes as easy as to query the prefix. + DHT maintenance mechanisms for both voluntary or ungraceful departures of peers. - Performance/accuracy trade-off in choosing the prefix length to be used. + Developed with mobile, ad-hoc networks in mind. + Evenly distributed IDs.
DynaMo - Winter et al. (2004)	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Random Landmarking (RLM). Closest Neighbor Prefix Assignment (CPNA). Making last hop as local as possible.	+ Dynamic landmarking is a fully distributed approach. - RLM may impose extra overhead to landmark nodes. - CPNA is a coarse-grained proximity approach. + Fine granularity of peer and object embedding (eight octants). + Load balance. - DHT with false negative queries. - Slow adaptation to network changes. + Continuously adaptive mechanism. + It can be considered lightweight. + Can coexist with other approaches (like landmarking). + Works sufficiently for large scopes as well as environments with high churn.
3DO - Abid et al. (2014)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	3 dimensional space.	+ Compared to Mithos, scales much better. - Needs synchronization within zones for every probing/jump for each node. - Contention situation in selective jump phase. - Probing phase incurs unnecessary traffic.
SAT-Match - Ren et al. (2004)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Selective jumps to adjust peer positioning in the DHT. Stretch reduction scheme.	

**Fig. 13.** Successive nodes in the overlay circle can be chosen to be from the same or near autonomous systems (AS).

discover peers in physical proximity. However, as already discussed, these methods do not always yield reliable estimates for node positions over the Internet. Landmark servers are not self-organizing and present maintenance overhead. Moreover, to serve a structured P2P-network with potentially tens of thousands or more peers, multiple such servers are required to be distributed over the world, which is a difficult proposition per se. RTTs can measure the de-

**Fig. 14.** Proximity routing takes proximity information into account during the routing process.

lay between peers, but their use constitutes a greedy method that can result in sub-optimal overlay topologies; the latter is especially true if nearby nodes happen to be connected through low-bandwidth connections.

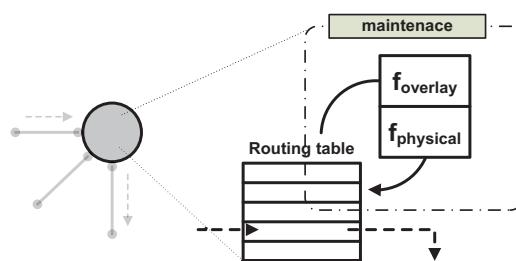


Fig. 15. Proximity neighbor selection dictates that proximity is taken into account during routing table maintenance.

1688 On one hand, approaches based on *geographic layout* improve the
1689 query efficiency of the system in general and are especially effective

when applied to CAN-like implementations. On the other hand, they
1690 inherently undermine the uniform distribution of peer identifiers.
1691 This is an important downside as it creates hot-spots that hurt sys-
1692 tem scalability. Moreover, it adversely affects system availability as
1693 nearby nodes are more likely to suffer collective failures (Huijin and
1694 Yongting, 2007).
1695

4.2.2. Proximity Routing (PR)

This category of approaches use physical proximity knowledge to
1696 decide the best next hop during routing. As shown in Fig. 14, main-
1697 tenance of routing tables themselves is a separate process based
1698 only on the system's DHT construction mechanism and overlay ID
1699 distribution.
1700

We reviewed an array of methodologies that strictly adhere to
1701 the tenets of this category such as *Mithos* and *PChord* as well as
1702

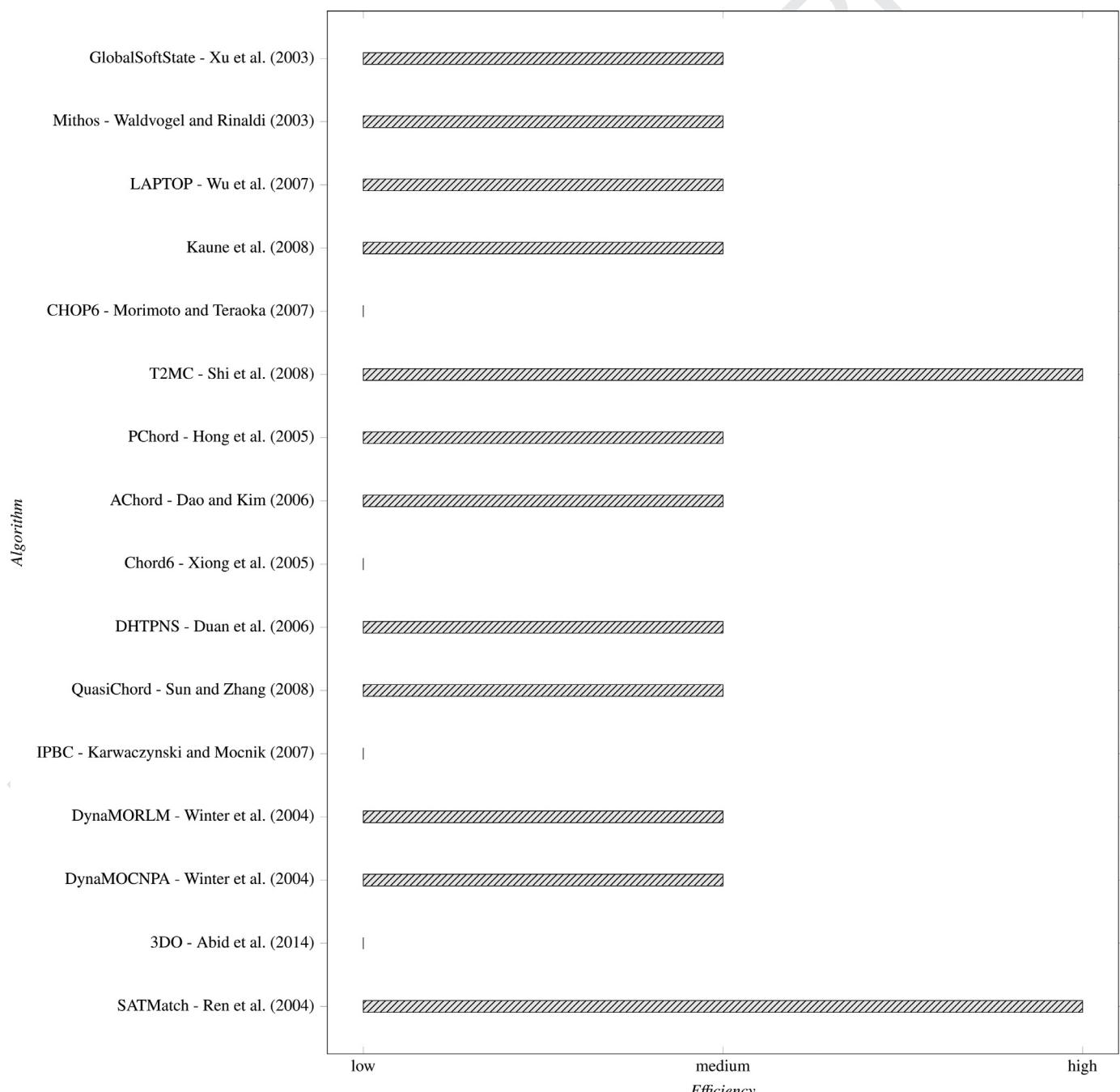


Fig. 16. Efficiency pictorial comparison of structured approaches.

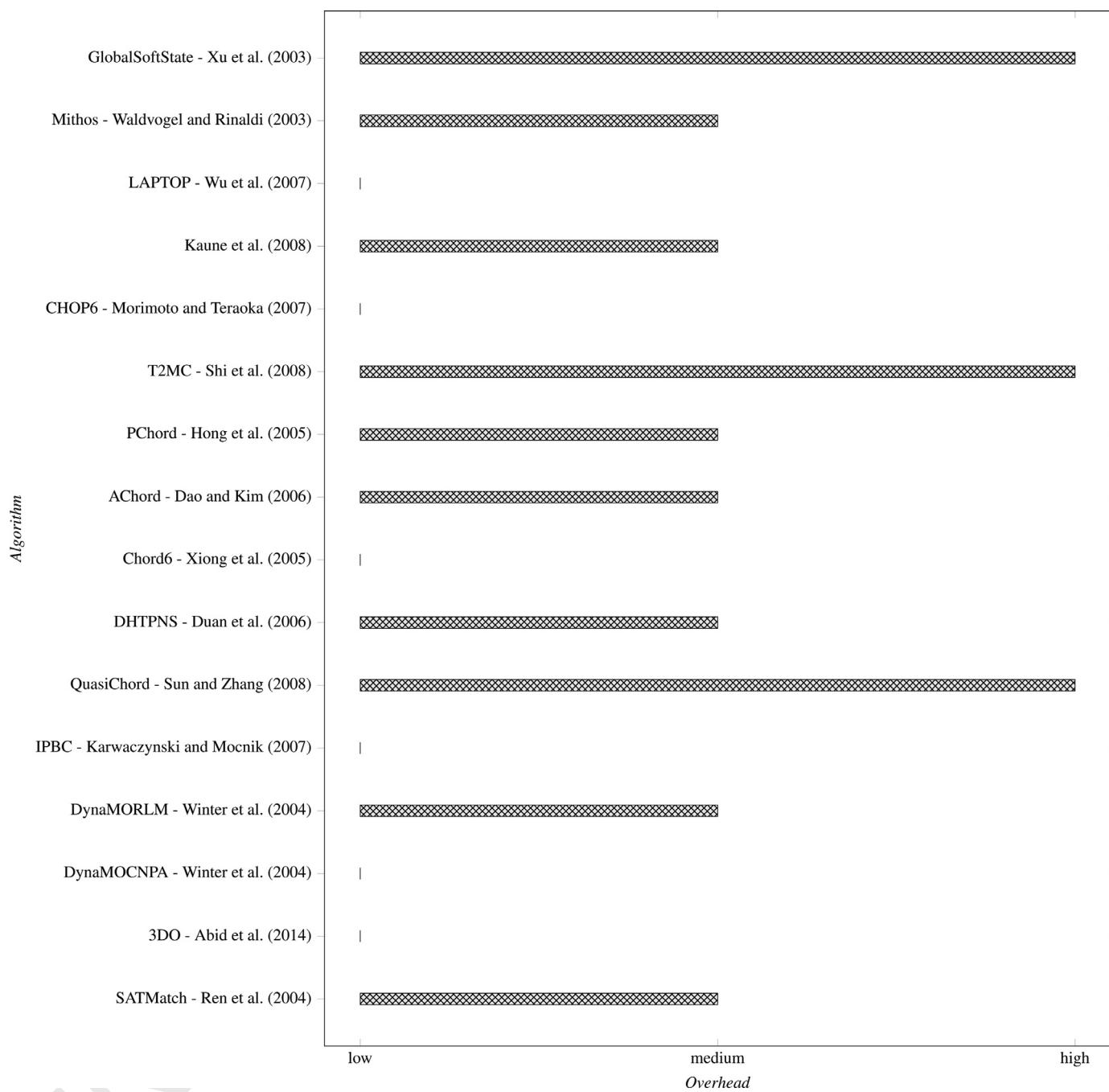


Fig. 17. Overhead pictorial comparison of structured approaches (lower is better).

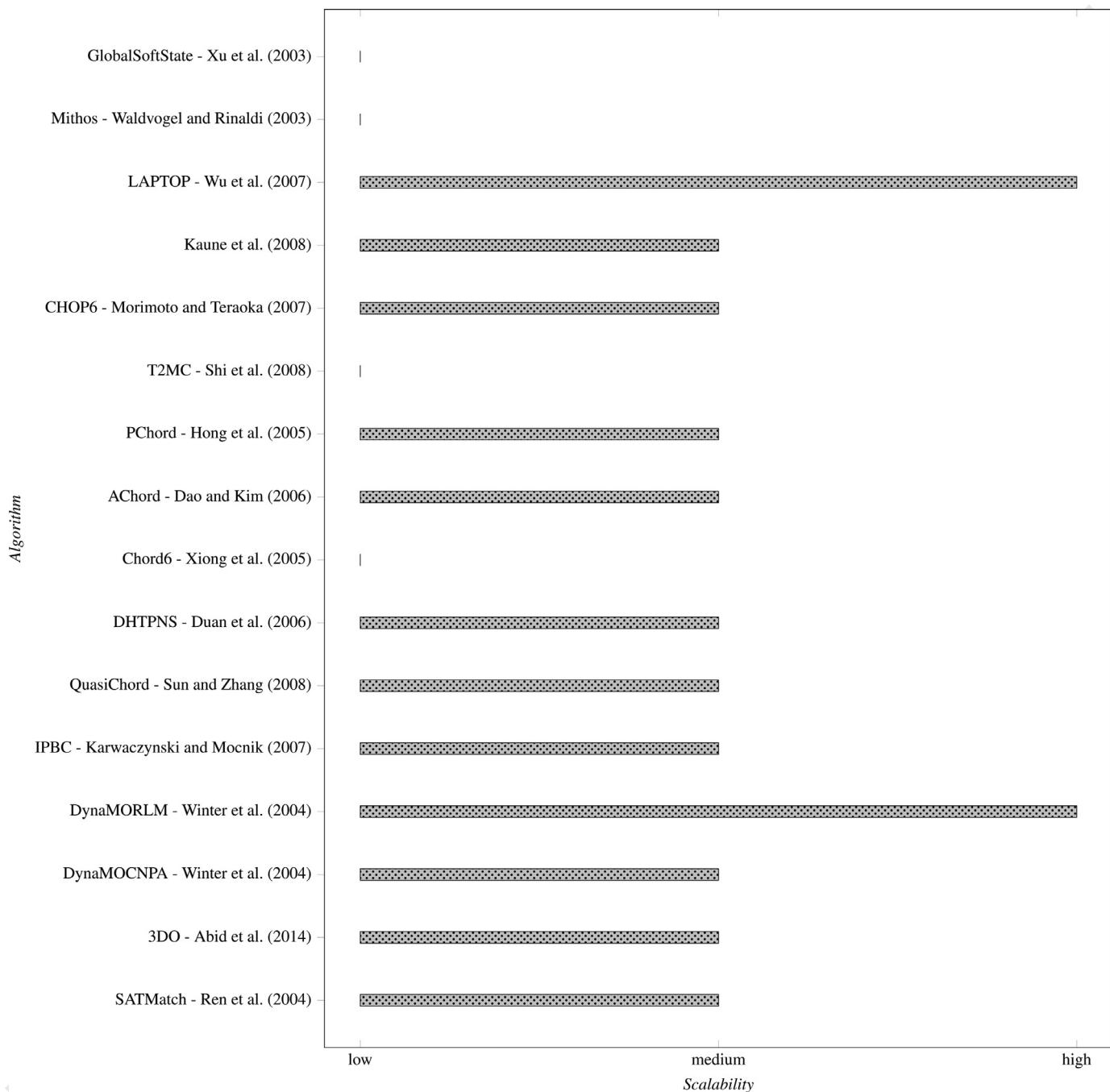
others that function along with proximity neighbor selection such as *DynaMO*. When *PR*-based algorithms route messages, they try to balance between choosing the node that will further progress the routing towards the destination and picking the closest entry in the routing table in terms of network proximity. In the degenerate case, a system could use a greedy approach to select only the lowest latency candidate peers during forwarding. This leads to longer, sub-optimal, paths to the destinations and thus imposes extra stress on the network. *PR* is also less effective than geographical layout when applied to *CAN*-like implementations.

4.2.3. Proximity Neighbor Selection (PNS)

Algorithms in this category use proximity data during routing table construction and maintenance along the lines of Fig. 15. The prox-

imity information used is different from that of landmark-based systems described in Section 4.2.1; here, RTT values between nodes (i.e., *SAT-Match*) or IP address prefixes (i.e., *DynaMO*) are predominantly used to detect proximity. Freedman et al. (2005) states that 97% of prefixes larger than 24 bits belong to a single geographical location. However, using a smaller number of bits creates less precise results and a larger number of bits may increase the burden in the network and simultaneously reduce the possible number of neighbors. Moreover, Hong et al. (2005) argue that prefix-based approaches can destroy the uniform distribution in the key space and do not work in one-dimensional key space where the mapping is restricted to operate at the overlay.

Therefore, a careful selection is required in terms of performance/accuracy trade-off. *Tapestry*, *Pastry*, and *CAN* successfully

**Fig. 18.** Scalability pictorial comparison of structured approaches.

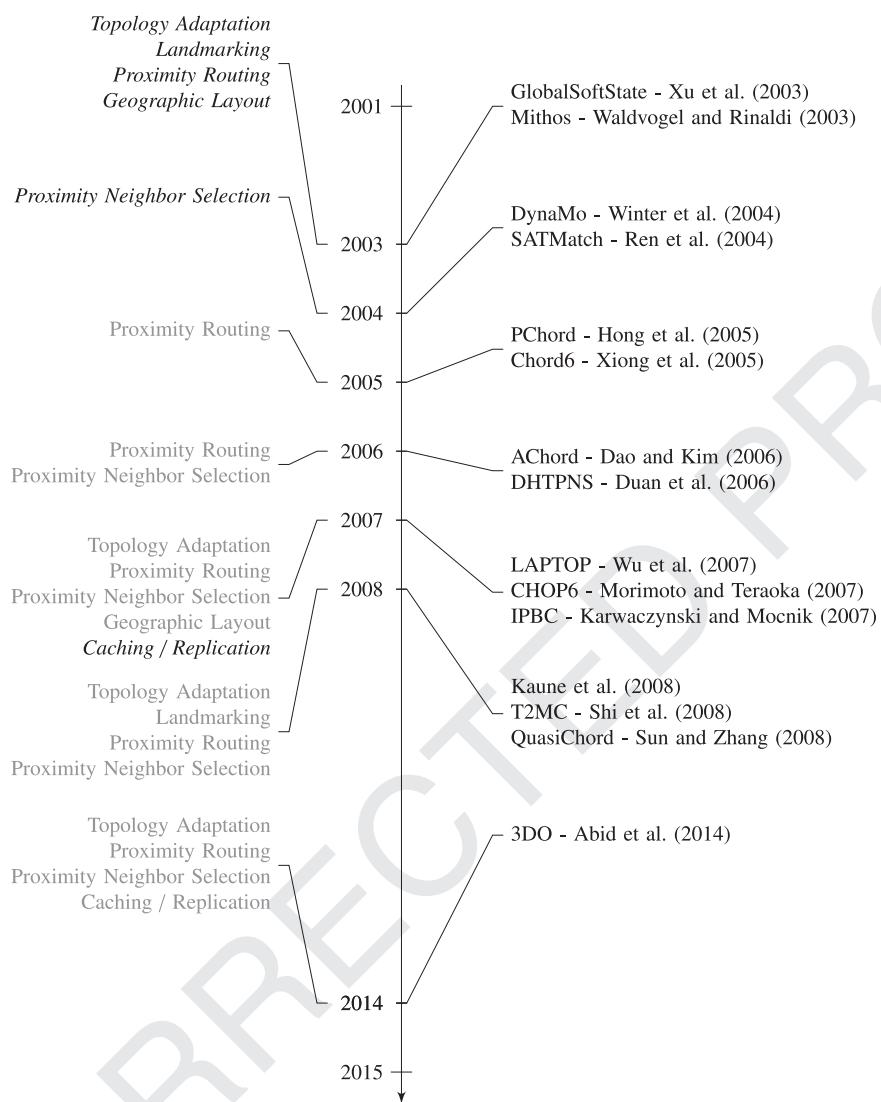


Fig. 19. Trends on characteristics of structured approaches (with emphasis denoting first appearance).

1731 incorporated proximity neighbor selection into their methods. The
 1732 routing protocol in *Pastry* is based on longest node ID prefix matching,
 1733 while *CAN* uses RTT values to detect nearby nodes. *Castro et al.*
 1734 (2002a) reports that proximity neighbor selection is an effective
 1735 proximity-based method. In general, proximity neighbor selection is
 1736 considered superior to proximity routing, however, simultaneous use
 1737 of both are also possible.

1738 5. Conclusion

1739 Development of P2P-systems and applications typically involve
 1740 juggling a set of trade-offs (e.g., choosing higher accuracy at the cost
 1741 of increased system overhead). In surveying more than a decade's
 1742 worth of research efforts aimed at solving the topology mismatch
 1743 problem in both unstructured and structured P2P networks, we find
 1744 this tussle-of-trade-offs property to hold. With regards to the three
 1745 criteria we used –efficiency, overhead, and scalability– we find that
 1746 none of the proposed solutions is superior to the others on all three
 1747 fronts. This is not surprising. Nonetheless, by (i) presenting an analy-
 1748 sis of each of the solutions, including what distinguishes it from oth-
 1749 ers as well as its advantages and disadvantages and (ii) offering a pic-
 1750 torial comparison of how each solution fares with regards to others
 1751 as far as efficiency, overhead, and scalability are concerned, we hope

1752 to provide P2P developers and researchers with enough insight and
 1753 perspective; as they face specific problems, they will be readily able
 1754 to draw the best possible design decisions for their applications.

1755 Acknowledgements

1756 We would like to thank the anonymous reviewers who helped
 1757 us improve the presentation of our work with their insightful com-
 1758 ments. This work has been partially supported by *iMarine FP7 EU*
 1759 project and *ERC Starting Grant no. 279237*.

1760 References

- Abid, S., Othman, M., Shah, N., 2014. 3D P2P overlay over MANETs. *Comput. Netw.* 64, 89–111. doi:[10.1016/j.comnet.2014.02.006](https://doi.org/10.1016/j.comnet.2014.02.006). <http://www.sciencedirect.com/science/article/pii/S1389128614000486>.
- Adar, E., Huberman, B.A., 2000. Free riding on Gnutella. *First Monday* 5, 2000.
- Adya, A., Bolosky, W., Castro, M., Chaiken, R., Cermak, G., Douceur, J., Howell, J., Lorch, J., Theimer, M., Wattenhofer, R., 2002. FARSITE: Federated, Available, and Reliable Storage for an Incompletely Trusted Environment. Proceedings of the 5th Symposium on Operating Systems Design and Implementation (OSDI). USENIX, Boston, MA.
- Alima, L.O., Ghodsi, A., El-Ansary, S., Brand, P., Haridi, S., 2003. Multicast in DKS(N, k, f) Overlay Networks. Proceedings of the 3rd International Conference on Peer-to-Peer Computing (P2P). IEEE, Linkoping, Sweden.
- Bayardo Jr., R.J., Agrawal, R., Gruhl, D., Somani, A., 2002. YouServ: a Web-hosting and Content Sharing Tool for the Masses. Proceedings of the 11th International Conference on World Wide Web (WWW). W3C, Honolulu, HI.

- 1776 Bernstein, D., Feng, Z., Levine, B.N., Zilberstein, S., 2003. Adaptive Peer Selection. Proceedings of the 2nd International Workshop on Peer-to-Peer Systems (IPTPS '03).
1777 NSF, Berkeley, CA.
- 1778 Bindal, R., Cao, P., Chan, W., Medved, J., Suwala, G., Bates, T., Zhang, A., 2006. Improving Traffic Locality in BitTorrent via Biased Neighbor Selection. Proceedings of the 26th IEEE Int. Conf. on Distributed Computing Systems (ICDCS). IEEE, Lisbon, Portugal.
- 1783 BitTorrent, I., 2012. BitTorrent and μTorrent Software Surpass 150 Million User Milestone; Announce New Consumer Electronics Partnerships. http://www.bittorrent.com/intl/es/company/about/ces_2012_150m_users. On-line; accessed April 24th, 2015.
- 1787 Bolosky, W.J., Douceur, J.R., Ely, D., Theimer, M., 2000. Feasibility of a serverless distributed file system deployed on an existing set of desktop PCs. Perform. Eval. Rev. 28, 34–43. ACM Sigmetrics.
- 1790 Buford, J.F., Yu, H., Liao, E.K., 2009. P2P Networking and Applications. Morgan Kaufmann, Burlington, MA.
- 1792 Cai, H., Wang, J., 2004. Foreseer: a Novel, Locality-aware Peer-to-peer System Architecture for Keyword Searches. Proceedings of the 5th ACM/FIP/USENIX International Conference on Middleware. Springer-Verlag Inc., Toronto, Canada, pp. 38–58. <http://dl.acm.org/citation.cfm?id=1045658.1045663>.
- 1795 Castro, M., Druschel, P., Charlie, Y., Rowstron, H.A., 2002a. Exploiting Network Proximity in Peer-to-Peer Overlay Networks. Technical report, Microsoft Research, MSR-TR-2002-82.
- 1799 Castro, M., Druschel, P., Hu, Y.C., Rowstron, A., 2002b. Exploiting Network Proximity in Distributed Hash Tables. Proceedings of International Workshop on Future Directions in Distributed Computing (FuDiCo). Springer-Verlag, Bertinoro, Italy, pp. 52–55.
- 1803 Castro, M., Druschel, P., Hu, Y.C., Rowstron, A., 2003. Topology-aware Routing in Structured Peer-to-Peer Overlay Networks. Proceedings of Future Directions in Distributed Computing (FuDiCo). Springer-Verlag, Bertinoro, Italy.
- 1806 Castro, M., Druschel, P., Kermarrec, A.M., Rowstron, A.I.T., 2002c. Scribe: a large-scale and decentralized application-level multicast infrastructure. IEEE J. Sel. Areas in Commun. 20, 1489–1499.
- 1809 Chandra, J., Shaw, S., Ganguly, N., 2010. HPC5: an efficient topology generation mechanism for Gnutella networks. Comput. Netw. 54, 1440–1459. <http://dx.doi.org/10.1016/j.comnet.2009.11.017>.
- 1812 Chawathe, Y., 2000. An Architecture for Internet Content Distribution as an Infrastructure Service, Ph.D. thesis. University of California, Berkeley, CA. <http://research.chawathe.com/people/yatin/publications/docs/thesis-single.ps.gz>.
- 1815 Chawathe, Y., Ratnasamy, S., Breslau, L., Lanham, N., Shenker, S., 2003. Making Gnutella-like P2P Systems Scalable. Proceedings of the 2003 SIGCOMM Conference. ACM, Karlsruhe, Germany, pp. 407–418.
- 1818 Chen, X., Li, Z., Zhuang, Y., Han, J., Chen, L., 2006. HAND: An Overlay Optimization Algorithm in Peer-to-Peer Systems. In: Lecture Notes in Computer Science, 4208/2006. Springer, Berlin, Germany, pp. 290–299.
- 1821 Choffnes, D.R., Bustamante, F.E., 2008. Taming the torrent: a practical approach to reducing cross-ISP traffic in peer-to-peer systems. ACM SIGCOMM Comput. Commun. Rev. 38, 363–374.
- 1824 Chu, Y., Rao, S., Seshan, S., Zhang, H., 2001. Enabling conferencing applications on the internet using an overlay multicast architecture. ACM SIGCOMM Comput. Commun. Rev. 31, 55–67.
- 1827 Chu, Y., Rao, S., Seshan, S., Zhang, H., 2002. A case for end system multicast. IEEE J. Sel. Areas Commun. 20, 1456–1471.
- 1829 Chu, Y., Rao, S.G., Zhang, H., 2000. A Case for End System Multicast. Proceedings of the ACM SIGMETRICS International Conference. ACM, Santa Clara, CA, pp. 1–12.
- 1831 Cisco, 2014. White Paper: Cisco Visual Networking Index: Forecast and Methodology, 2013 – 2018. Technical Report. Cisco. http://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white-paper_c11-481360.pdf.
- 1835 Clarke, I., Sandberg, O., Wiley, B., Hong, T.W., 2000. Freenet: A Distributed Anonymous Information Storage and Retrieval System. Workshop on Design Issues in Anonymity and Unobservability. Springer, Berkeley, CA, pp. 46–66.
- 1838 Coates, M., Hero, A., Nowak, R., Yu, B., 2002. Internet tomography. IEEE Signal Process. Mag. 19, 47–65.
- 1840 Cohen, B., 2003. Incentives Build Robustness in BitTorrent. Workshop on Economics of Peer-to-Peer Systems. UC-Berkeley, Berkeley, CA.
- 1842 Cohen, E., Shenker, S., 2002. Replication Strategies in Unstructured Peer-to-Peer Networks. Proceedings of the ACM SIGCOMM Conference. ACM, Pittsburgh, PA, pp. 177–190.
- 1845 Cooper, B.F., 2005. Quickly Routing Searches Without Having to Move Content. Proceedings of the 4th International Conference on Peer-to-Peer Systems. Springer-Verlag, Ithaca, NY, pp. 163–172. http://dx.doi.org/10.1007/11558989_15.
- 1848 Costa, M., Castro, M., Rowstron, A., Key, P., 2004. PIC: Practical Internet Coordinates for Distance Estimation. Proceedings of 24th International Conference on Distributed Computing Systems (ICDCS'04). IEEE Computer Society, Washington, DC, pp. 178–187. <http://dl.acm.org/citation.cfm?id=977400.977998>.
- 1852 Cox, R., Dabek, F., Kaashoek, F., Li, J., Morris, R., 2004. Practical, distributed network coordinates. ACM SIGCOMM Comput. Commun. Rev. 34, 113–118.
- 1854 Cox, R., Muthitacharoen, A., Morris, R., 2002. Serving DNS using a peer-to-peer lookup service. In: Druschel, P., Kaashoek, F., Rowstron, A. (Eds.), Peer-to-Peer Systems. In: Lecture Notes in Computer Science, chapter 15, Vol. 2429. Springer Berlin / Heidelberg, Berlin, Heidelberg, pp. 155–165. http://dx.doi.org/10.1007/3-540-45748-8_15.
- 1856 Crespo, A., Garcia-Molina, H., 2002. Routing Indices for Peer-to-peer Systems. Proceedings of the 22nd International Conference on Distributed Computing Systems (ICDCS'02). IEEE Computer Society, Vienna, Austria, pp. 23–33. <http://dl.acm.org/citation.cfm?id=850928.851858>.
- 1858 Cuevas, R., Laotaris, N., Yang, X., Siganos, G., Rodriguez, P., 2009. Deep diving into Bit-Torrent locality. CoRR abs/0907.3874, 1–16.
- 1860 Dabek, F., Kaashoek, M.F., Karger, D., Morris, R., Stoica, I., 2001. Wide-area Cooperative Storage with CFS. Proceedings of the 18th ACM Symposium on Operating Systems Principles (SOSP). ACM, Banff, Canada, pp. 202–215.
- 1862 Dao, L.H., Kim, J., 2006. AChord: Topology-Aware Chord in Anycast-Enabled Networks. Proceedings of International Conference on Hybrid Information Technology (ICHIT). IEEE Computer Society, Jeju Island, Korea, pp. 334–341. <http://dx.doi.org/10.1109/ICHIT.2006.47>.
- 1864 Druschel, P., Rowstron, A., 2001. PAST: A Large-Scale, Persistent Peer-to-Peer Storage Utility. Proceedings of 8th Workshop on Hot Topics in Operating Systems (HotOS). IEEE Computer Society, Elmau/Oberbayern, Germany, pp. 75–80.
- 1866 Duan, H., Lu, X., Tang, H., Zhou, X., Zhao, Z., 2006. Proximity Neighbor Selection in Structured P2P Network. Proceedings of the 6th IEEE International Conference on Computer and Information Technology (CIT). IEEE Computer Society, Seoul, Korea, p. 52.
- 1868 Eom, H., Wolinsky, D., Figueiredo, R.J., 2011. SOLARE: Self-Organizing Latency-Aware Resource Ensemble. Proceedings of the 13th IEEE International Conference on High Performance Computing and Communications (HPCC). IEEE, Banff, Canada, pp. 229–236. doi: 10.1109/HPCC.2011.38.
- 1870 Ferguson, D., 2006. P2P File Sharing - The Evolving Distribution Chain. <http://retroshare.sourceforge.net>.
- 1872 Freedman, M.J., Vutukuru, M., Feamster, N., Balakrishnan, H., 2005. Geographic Locality of IP Prefixes. Proceedings of the 5th SIGCOMM Conference on Internet Measurement (IMC). USENIX, Berkeley, CA, 13–13.
- 1874 Garey, M.R., Johnson, D.S., 1979. Computers and Intractability: A Guide to the Theory of NP-Completeness. Books in Mathematical Sciences. W. H. Freeman and Co, New York, NY.
- 1876 Gnutella, 2000. The Gnutella Protocol Specification v0.4. http://www9.limewire.com/developer/gnutella_protocol0.4.pdf.
- 1878 Gnutella, 2009. The Gnutella Protocol Specification v0.6. <http://rfc-gnutella.sourceforge.net/>. On-line; accessed August 15th, 2014.
- 1880 Grönvall, B., Westerlund, A., Danielsson, J., 2010. Arla Project. <http://www.stacken.kth.se/project/arla>. On-line; accessed on April 20th, 2015.
- 1882 Gummadi, K., Saroiu, S., Gribble, S.D., 2002. King: Estimating Latency Between Arbitrary Internet end Hosts. Proceedings of the 2nd ACM SIGCOMM Workshop on Internet Measurement. ACM, Marseille, France, pp. 5–18. <http://doi.acm.org/10.1145/637201.637203>.
- 1884 Hong, F., Li, M., Yu, J., Wang, Y., 2005. PChord: Improvement on Chord to Achieve Better Routing Efficiency by Exploiting Proximity. Proceedings of the 1st International Workshop on Mobility in Peer-to-Peer Systems (MPPS). IEEE Computer Society, Columbus, OH, pp. 806–811. <http://dx.doi.org/10.1109/ICDCSW.2005.108>.
- 1886 Hsiao, H.C., Liao, H., Huang, C.C., 2009. Resolving the topology mismatch problem in unstructured peer-to-peer networks. IEEE Trans. Parallel and Distrib. Syst. 20, 1668–1681. <http://doi.ieeecomputersociety.org/10.1109/TPDS.2009.24>.
- 1888 Hsiao, H.C., Liao, H., Yeh, P.S., 2010. A near-optimal algorithm attacking the topology mismatch problem in unstructured peer-to-peer networks. IEEE Trans. Parallel Distrib. Syst. 21, 983–997. doi: 10.1109/TPDS.2009.160.
- 1890 Hughes, D., Coulson, G., Walkerline, J., 2005. Free riding on Gnutella revisited: the bell tolls? IEEE Distrib. Syst. On-line 6, 1–13.
- 1892 Hughes, D., Walkerline, J., Lee, K., 2008. Socio-Technical Issues in P2P File Sharing Systems. <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.62.1347>. On-line; accessed in January 2015.
- 1894 Huijin, W., Yongting, L., 2007. Cone: A Topology-Aware Structured P2P System with Proximity Neighbor Selection. Proceedings of Future Generation Communication and Networking (FGCN). IEEE Computer Society, Jeju-Island, Korea, pp. 43–49.
- 1896 Iyer, S., Rowstron, A., Druschel, P., 2002. Squirrel: A Decentralized Peer-to-Peer Web Cache. Proceedings of 21st Annual Symposium on Principles of Distributed Computing (PODC). ACM, Monterey, CA.
- 1898 Jiang, S., Guo, L., Zhang, X., Wang, H., 2008. LightFlood: minimizing redundant messages and maximizing scope of peer-to-peer search. IEEE Trans. Parallel and Distrib. Syst. 19, 601–614. <http://dx.doi.org/10.1109/TPDS.2007.70772>.
- 1900 JupiterMediaMetrix, 2001. Global Napster Usage Plummetts, But New File-Sharing Alternatives Gaining Ground, Press Release. <http://web.archive.org/web/20080413104420/http://www.comscore.com/press/release.asp?id=249>. on-line; accessed October 9th, 2014.
- 1902 Karagiannis, T., Rodriguez, P., Papagiannaki, K., 2005. Should Internet Service Providers Fear Peer-assisted Content Distribution? Proceedings of the 5th ACM SIGCOMM Conference on Internet Measurement (ICM). USENIX Association, Berkeley, CA, p. 6. <http://portal.acm.org/citation.cfm?id=1251092>.
- 1904 Karp, B., Kung, H.T., 2000. GPRS: Greedy Perimeter Stateless Routing for Wireless Networks. Proceedings of the 6th annual international conference on Mobile computing and networking. ACM, New York, NY, USA, pp. 243–254. <http://dx.doi.org/10.1145/345910.345953>.
- 1906 Karwaczynski, P., Mocnik, J., 2007. IP-based clustering for peer-to-peer overlays. J. Softw. 2, 30–37.
- 1908 Kaune, S., Lauinger, T., Kovacevic, A., Pussep, K., 2008. Embracing the Peer Next Door: Proximity in Kademlia. Proceedings of the 8th International Conference on Peer-to-Peer Computing (P2PC). IEEE, Aachen, Germany, pp. 343–350.
- 1910 Kazaa, 2009. <http://www.kazaa.com/>. <http://www.kazaa.com/>.

- 1943 Kubiakowicz, J., Bindel, D., Chen, Y., Eaton, P., Geels, D., Gummadi, R., Rhea, S., Weatherspoon, H., Weimer, W., Wells, C., Zhao, B., 2000. OceanStore: An Architecture for Global-scale Persistent Storage. Proceedings of the ACM ASPLOS-IX. ACM, Cambridge, MA.
- 1947 Li, W., Chen, S., Yu, T., 2008. UTAPS: An Underlying Topology-Aware Peer Selection Algorithm in BitTorrent. Proceedings of the 22nd International Conference on Advanced Information Networking and Applications (AINA). IEEE Computer Society, Okinawa, Japan, pp. 539–545. <http://dx.doi.org/10.1109/AINA.2008.92>.
- 1951 Liu, B., Cui, Y., Lu, Y., Xue, Y., 2009. Locality-Awareness in BitTorrent-like P2P Applications. *IEEE Trans. Multimed.* 11, 361–171.
- 1953 Liu, Y., 2008. A two-hop solution to solving topology mismatch. *IEEE Trans. Parallel Distrib. Syst.* 19, 1591–1600.
- 1955 Liu, Y., Liu, X., Xiao, L., Ni, L., Zhang, X., 2004a. Location-Aware Topology Matching in P2P Systems. Proceedings of the 23rd IEEE INFOCOM Conference. IEEE, Hong Kong, China, pp. 2220–2230.
- 1958 Liu, Y., Ni, L.M., Xiao, L., Esfahanian, A.H., 2005a. Approaching Optimal Peer-to-Peer Overlays. Proceedings of the 13th IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MAS-COTS). IEEE Computer Society, Atlanta, GA, pp. 407–414. <http://dx.doi.org/10.1109/MASCOT.2005.15>.
- 1963 Liu, Y., Xiao, L., Liu, X., Ni, L.M., Zhang, X., 2005b. Location awareness in unstructured peer-to-peer systems. *IEEE Trans. Parallel and Distrib. Syst.* 16, 163–174. <http://dx.doi.org/10.1109/TPDS.2005.21>.
- 1966 Liu, Y., Xiao, L., Ni, L., 2007. Building a scalable bipartite P2P overlay network. *IEEE Trans. Parallel Distrib. Syst.* 18, 1296–1306. <http://dx.doi.org/10.1109/TPDS.2007.1059>.
- 1968 Liu, Y., Zhuang, Z., Xiao, L., Ni, L., 2003. AOTO: Adaptive Overlay Topology Optimization in Unstructured P2P Systems. Proceedings of the 2003 IEEE GLOBECOM Conference. IEEE, San Francisco, CA, pp. 4186–4190. doi:[10.1109/GLOCOM.2003.1259015](https://doi.org/10.1109/GLOCOM.2003.1259015).
- 1972 Liu, Y., Zhuang, Z., Xiao, L., Ni, L.M., 2004b. A Distributed Approach to Solving Overlay Mismatching Problem. Proceeding of the 24th International Conference on Distributed Computing Systems (ICDCS). IEEE Computer Society, Tokyo, Japan, pp. 132–139.
- 1976 Lv, Q., Cao, P., Cohen, E., Li, K., Shenker, S., 2002. Search and Replication in Unstructured Peer-to-peer Networks. Proceedings of the 16th International Conference on Supercomputing (ICS). ACM, New York, NY, pp. 84–95. <http://doi.acm.org/10.1145/514191.5154206>.
- 1980 Manku, G., 2003. Routing Networks for Distributed Hash Tables. Proceedings of the 22nd ACM Symposium on Principles of Distributed Computing. ACM, Boston, MA, pp. 133–142. <http://doi.acm.org/10.1145/872035.872054>.
- 1983 Maymounkov, P., Mazières, D., 2002. Kademlia: A Peer-to-Peer Information System Based on the XOR Metric. Proceedings of the 1st International Workshop on Peer-to-Peer Systems (IPTPS). Springer-Verlag, London, UK, pp. 53–65.
- 1986 Mercier, J., 2013. 50 Million Concurrent Users On-line! skypenumerology.blogspot.se/2013/01/50-million-concurrent-users-on-line.html. On-line; accessed on April, 19th 2015.
- 1989 Metz, C., 2002. IP Anycast: point-to-(Any) point communication. *IEEE Internet Comput.* 6, 94–98.
- 1991 Microsoft, 2012. Three Degrees. <http://web.archive.org/web/20040103063458/http://threedeegrees.com/>. archive; accessed on March 14th, 2015.
- 1993 Mondéjar, R., García-López, P., Pariot, C., Pamies-Juarez, L., 2013. CloudSNAP : a transparent infrastructure for decentralized web deployment using distributed interception. *Future Gener. Comput. Syst.* 29, 370–380.
- 1996 Morimoto, S., Teraoka, F., 2007. Chop6: A dht routing mechanism considering proximity. Proceedings of the 2007 International Symposium on Applications and the Internet Workshops (SAINT-W). IEEE Comput. Soc., Hiroshima, Japan, p. 59.
- 1999 Muthitacharoen, A., Morris, R., Gil, T.M., Chen, B., 2002. Ivy: A Read/Write Peer-to-Peer File System. Proceedings of 5th Symposium on Operating Systems Design and Implementation (OSDI). USENIX Association, Boston, MA.
- 2002 Napster, s., 2007. <http://en.wikipedia.org/wiki/Napster>. On-line; accessed on April 20th, 2015.
- 2004 Ng, T.S.E., Zhang, H., 2001. Towards Global Network Positioning. Proc. of the 1st ACM-SIGCOMM Workshop on Internet Measurement (IMW). ACM, San Francisco, CA, pp. 25–29. <http://doi.acm.org/10.1145/505202.505206>.
- 2007 Overnet, 2009. <http://www.overnet.org/>. <http://www.overnet.org/>.
- 2008 Papadakis, H., Fragopoulou, P., Markatos, E., Roussopoulos, M., 2013. ITA: Innocuous Topology Awareness for unstructured P2P networks. *IEEE Transactions on Parallel and Distrib. Syst.* 24, 1589–1601. doi:[10.1109/TPDS.2012.137](https://doi.org/10.1109/TPDS.2012.137).
- 2011 Papadimitriou, A., Delis, A., 2008. Data Dissemination in Unstructured Peer-to-Peer Networks. Proceedings of the 37th International Conference on Parallel Processing (ICPP'08). IEEE, Portland, OR.
- 2014 Piatek, M., Madhyastha, H.V., John, J., Krishnamurthy, A., Anderson, T., 2009. Pitfalls for ISP-friendly P2P Design. Proceedings of the 8th ACM Workshop on Hot Topics in Networks (HotNets-VIII). ACM, New York, NY.
- 2017 YaCY Project, 2012. <http://yacyweb.de>. on-line; accessed on March 14th, 2015.
- 2018 Pyun, Y.J., Reeves, D.S., 2004. Constructing a Balanced, $(\log(N)/\log(\log(N))$ -Diameter Super-Peer Topology for Scalable P2P Systems. Proceedings of the IEEE International Conference on Peer-to-Peer Computing. IEEE Computer Society, Zurich, Switzerland, pp. 210–218. <http://doi.ieeecomputersociety.org/10.1109/PTP.2004.1334949>.
- 2023 Qin, F., Liu, J., Zheng, L., Ge, L., 2009. An Effective Network-Aware Peer Selection Algorithm in BitTorrent. Proceedings of 5th International Conference on Intelligent Information Hiding and Multimedia Signal Processing. IEEE Computer Society, Kyoto, Japan, pp. 418–421. <http://dx.doi.org/10.1109/IH-MSP.2009.90>.
- 2027 Qiu, T., Chen, G., Ye, M., Chan, E., Zhao, B.Y., 2007. Towards Location-aware Topology in both Unstructured and Structured P2P Systems. Proceedings of the 2007 International Conference on Parallel Processing (ICPP). IEEE Computer Society, Xian, China. <http://dx.doi.org/10.1109/ICPP.2007.81>.
- 2029 Ramasubramanian, V., Sirer, E.G., 2004. Beehive: O(1)lookup performance for power-law query distributions in peer-to-peer overlays. Proceedings of the 1st Symposium on Networked Systems Design and Implementation - Volume 1. USENIX Association, San Francisco, CA. <http://dl.acm.org/citation.cfm?id=1251175.1251183>, pages 8–8
- 2030 Ratnasamy, S., Francis, P., Handley, M., Karp, R., Shenker, S., 2001. A Scalable Content-Addressable Network. Proceedings of the 2001 ACM SIGCOMM Conference. ACM, San Diego, CA, pp. 161–172. <http://doi.acm.org/10.1145/383059.383072>.
- 2031 Ratnasamy, S., Handley, M., Karp, R., Shenker, S., 2002a. Topologically-Aware Overlay Construction and Server Selection. Proceedings of the 21st IEEE INFOCOM. IEEE, New York, NY, pp. 1190–1199.
- 2032 Ratnasamy, S., Stoica, I., Shenker, S., 2002b. Routing Algorithms for DHTs: Some Open Questions. In: Proceedings of 1st International Workshop on Peer-to-Peer Systems (IPTPS). In: LNCS-2429. Springer-Verlag, Cambridge, MA, pp. 45–52.
- 2033 Ren, S., Guo, L., Jiang, S., Zhang, X., 2004. SAT-Match: A Self-Adaptive Topology Matching Method to Achieve Low Lookup Latency in Structured P2P Overlay Networks. Proceedings of IEEE Parallel and Distributed Processing Symposium (IPDPS). IEEE Computer Society, Santa Fe, NM, pp. 83–92. <http://doi.ieeecomputersociety.org/10.1109/IPDPS.2004.1303022>.
- 2034 Ren, S., Tan, E., Luo, T., Chen, S., Guo, L., Zhang, X., 2010. TopBT: A Topology-Aware and Infrastructure-Independent BitTorrent Client. Proceedings of 2010 IEEE INFOCOM. IEEE, San Diego, CA.
- 2035 Rhea, S., Geels, D., Roscoe, T., Kubiakowicz, J., 2004. Handling Churn in a DHT. Proceedings of the USENIX Annual Technical Conference (ATC). USENIX Association, Boston, MA. <http://citeseerk.ist.psu.edu/viewdoc/summary?doi=10.1.1.10.9928>.
- 2036 Rinaldi, R., 2002. Routing and data location in overlay peer-to-peer networks. Institut Eurécom and Università degli Studi di Milano Diploma thesis. Also available as IBM Research Report RZ-3433.
- 2037 Ripeanu, M., Iamnitchi, A., Foster, I., 2002. Mapping the Gnutella network. *IEEE Internet Comput.* 6, 50–57. doi:[10.1109/4236.978369](https://doi.org/10.1109/4236.978369).
- 2038 Ritter, J., 2001. Why Gnutella Can't Scale. No, Really. <http://www.darkridge.com/~jpr/doc/gnutella.html>. On-line; accessed on November 30th, 2014.
- 2039 Rowstron, A.I.T., Druschel, P., 2001. Pastry: Scalable, Decentralized Object Location, and Routing for Large-Scale Peer-to-Peer Systems. Proceedings of the IFIP/ACM International Conference on Distributed Systems Platforms (Middleware). Springer-Verlag, Heidelberg, Germany, pp. 329–350.
- 2040 Rowstron, A.I.T., Kermarrec, A.M., Castro, M., Druschel, P., 2001. SCRIBE: The Design of a Large-Scale Event Notification Infrastructure. Proceedings of the 3rd International Workshop in Networked Group Communications (NGC). LCCS 2233, Springer, London, UK, pp. 30–43.
- 2041 Saroiu, S., Gummadi, K.P., Dunn, R.J., Gribble, S.D., Levy, H.M., 2002a. An Analysis of Internet Content Delivery Systems. Proceedings of 5th Symposium on Operating Systems Design and Implementation (OSDI). ACM, Boston, MA, pp. 315–327. <http://doi.acm.org/10.1145/1060289.1060319>.
- 2042 Saroiu, S., Gummadi, P.K., Gribble, S.D., 2002b. A Measurement Study of Peer-to-Peer File Sharing Systems. Vol. 4673. Proceedings of SPIE Multimedia Computing and Networking (MMCN). SPIE, San Jose, CA. citeseerk.nj.nec.com/saroiu02measurement.html.
- 2043 Schulze, H., Mochalski, K., 2009. Ipoque: Internet study 2008/2009. <https://www.christopher-parsons.com/Main/wp-content/uploads/2009/04/ipoque-internet-study-08-09.pdf>.
- 2044 Sen, S., Wang, J., 2004. Analyzing peer-to-peer traffic across large networks. *IEEE/ACM Trans. Netw.* 12, 219–232. doi:[10.1109/TNET.2004.826277](https://doi.org/10.1109/TNET.2004.826277).
- 2045 Shi, G., Long, Y., Chen, J., Gong, H., Zhang, H., 2008. T2MC: A Peer-to-Peer Mismatch Reduction Technique by Traceroute and 2-Means Classification Algorithm. Proceedings of the 2008 International Conference on NETWORKING: Ad Hoc and Sensor Networks, Wireless Networks, Next Generation Internet. IFIP Springer, LNCS 4982, Singapore, pp. 366–374.
- 2046 Skype, s., 2012. <http://www.skype.com>. On-line; accessed on April 24th, 2015.
- 2047 Smaragdakis, G., Lekakis, V., Laoutaris, N., Bestavros, A., Byers, J.W., Roussopoulos, M., 2008. The EGOIST Overlay Routing System. Proceedings of ACM CoNEXT Conference. ACM, Madrid, Spain.
- 2048 Srivatsa, M., Gedik, B., Liu, L., 2006. Large scaling unstructured peer-to-peer networks with heterogeneity-aware topology and routing. *IEEE Trans. on Parallel and Distrib. Syst.* 17, 1277–1293. <http://dx.doi.org/10.1109/TPDS.2006.158>.
- 2049 Stoica, I., Morris, R., Karger, D., Kaashoek, M.F., Balakrishnan, H., 2001. Chord: A Scalable Peer-to-Peer Lookup Service for Internet Applications. Proceedings of 2001 ACM SIGCOMM Conference. ACM, San Diego, CA, pp. 149–160. <http://doi.acm.org/10.1145/383059.383071>.
- 2050 Stutzbach, D., Rejaie, R., 2006. Understanding Churn in Peer-to-Peer Networks. Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement (IMC). ACM, Rio de Janeiro, Brazil, pp. 189–202. <http://doi.acm.org/10.1145/1177080.1177105>.
- 2051 Sun, M., Zhang, Z., 2008. Quasi-Chord: Physical Topology Aware Structured P2P Network. Proceedings of the 11th Joint Conference on Information Sciences (JCIS-2008). Atlantis Press, Shenzhen, China doi:[10.2991/jcis.2008.109](https://doi.org/10.2991/jcis.2008.109).
- 2052 TELECOMPAPER BV., 2011. Skype grows FY Revenues 20%, reaches 663 mln Users. <http://www.telecompaper.com/news/skype-grows-fy-revenues-20-reaches-663-mln-users>. On-line; accessed March 18th, 2015.
- 2053 TeleGeography/PriMetrica, I., 2014. Skype traffic continues to thrive. <https://www.telegeography.com/products/commsupdate/articles/2014/01/15/skype-traffic-continues-to-thrive/>. On-line; accessed on April, 20th 2015.
- 2054 Waldman, M., Mazières, D., 2001. Tangler: a Censorship-Resistant Publishing System Based on Document Entanglements. Proceedings of 8th ACM Conference on

- Computer and Communications Security (CCS). ACM, Philadelphia, PA, pp. 126–135. <http://dx.doi.org/10.1145/501983.502002>.
- Waldman, M., Rubin, A., Cranor, L., 2000. Publius: A Robust, Tamper-evident, Censorship-Resistant Web Publishing System. Proceedings 9th USENIX Security Symposium (SSYM). USENIX Association, Berkeley, CA, pp. 59–72.
- Waldbogel, M., Rinaldi, R., 2003. Efficient topology-aware overlay network. ACM SIGCOMM Comput. Commun. Rev. 33, 101–106.
- Winter, R., Zahn, T., Schiller, J., 2004. DynaMO: a topology-aware P2P overlay network for dynamic, mobile ad-hoc environments. Telecommun. Syst. 27, 321–345. doi: <10.1023/B:TELS.0000041014.05554.ee>.
- Wong, B., Slivkins, A., Sizer, E.G., 2005. Meridian: A Lightweight Network Location Service Without Virtual Coordinates. Proceedings of the 2005 SIGCOMM Conference. ACM, Philadelphia, PA, pp. 85–96. <http://doi.acm.org/10.1145/1080091.1080103>.
- Wu, C.J., Liu, D.K., Hwang, R.H., 2007. A location-aware peer-to-peer overlay network. Int. J. Commun. Syst. 20, 83–102. <http://dx.doi.org/10.1002/dac.v20:1>.
- Xing-feng, L., Bao-ping, Y., Wan-ming, L., 2008. Overlay multicast network optimization and simulation based on narada protocol. In: 10th International Conference on Advanced Communication Technology, 2008. ICACT 2008., vol. 3, pp. 2215–2220. doi: <10.1109/ICACT.2008.4494230>.
- Xiong, J., Zhang, Y., Hong, P., Li, J., 2005. Chord6: IPv6 Based Topology-Aware Chord. Proceedings the Joint International Conference on Autonomic and Autonomous Systems and International Conference on Networking and Services (ICAS/ICNS). IEEE Computer Society, Papete, Tahiti, p. 4.
- Xu, Z., Tang, C., Zhang, Z., 2003. Building Topology-Aware Overlays Using Global Soft-State. ICDCS '03: Proceedings of the 23rd International Conference on Distributed Computing Systems. IEEE Computer Society, Washington, DC, p. 500.
- Xu, Z., Zhang, Z., 2002. Building Low-maintenance Expressways for P2P Systems. Technical report. HP Laboratories Palo Alto. HP Laboratories, Internet Systems and Storage Laboratory.
- Yang, B., Garcia-Molina, H., 2002. Improving Search in Peer-to-Peer Networks. Proceedings of the 22d International Conference on Distributed Computing Systems (ICDCS). IEEE Computer Society, Vienna, Austria, pp. 5–14.
- Zeinalipour-Yazti, D., Kalogeraki, V., 2006. Structuring topologically aware overlay networks using domain names. Comput. Netw. 50, 3064–3082. <http://dx.doi.org/10.1016/j.comnet.2005.12.003>.
- Zeng, W., Zhu, Y., Lu, H., Zhuang, S., 2009. Path-diversity P2P overlay retransmission for reliable IP-multicast. IEEE Trans. Multimed. 11, 960–971. doi: <10.1109/TMM.2009.2021712>.
- Zhang, D., Lin, C., 2005. Efficient Delay Aware Peer-to-Peer Overlay Network. Advances in Web-Age Information Management. Springer, Berlin/Heidelberg, Germany, pp. 682–687. LNCS 3739.
- Zhang, X.Y., Zhang, Q., Zhang, Z., Song, G., Zhu, W., 2004. A construction of locality-aware overlay network: moverlay and its performance. IEEE J. Sel. Areas in Commun. 22, 18–28. doi: <10.1109/JSAC.2003.818780>.
- Zhao, B.Y., Kubiatowicz, J.D., Joseph, A.D., 2001. Tapestry: An Infrastructure for Fault-tolerant Wide-area Location and Routing. Technical report. University of California-Berkeley, Berkeley, CA. TR No. UCB/CSD-01-1141.
- Zhao, J., Lu, J., 2006. Solving overlay mismatching of unstructured p2p networks using physical locality information. P2P '06: Proceedings of the Sixth IEEE International Conference on Peer-to-Peer Computing. IEEE Comput. Soc., Washington, DC, pp. 75–76. <http://dx.doi.org/10.1109/P2P.2006.41>.
- Zheng, H., Lua, E.K., Pias, M., Griffin, T.G., 2005. Internet Routing Policies and Round-Trip-Times. Springer, LNCS 3431/2005, Berlin/Heidelberg, Germany, pp. 236–250. doi: <10.1007/b135479>.
- Zhu, Z., Kalnis, P., Bakiras, S., 2008. Dcmp: a distributed cycle minimization protocol for peer-to-peer networks. IEEE Trans. Parallel Distrib. Syst. 19, 363–377. doi: <10.1109/TPDS.2007.70732>.

Vassilis Moustakas received his MS Computer Science from the University of Athens and his BS in Informatics from the Technological Educational Institute of Athens. He is currently a member of the technical staff at the Amazon Development Center Germany in Munich as part of the A9.com Product Search group implementing scalable search and discovery technologies for Amazon.com and other websites. Prior to that, he worked as a software engineer for a number of companies as well as a researcher for various FP7 EU projects. His interests lie in Information Retrieval, Data Mining and Distributed Computing.

Hüseyin Akcan is an Associate Professor in the Department of Software Engineering at Izmir University of Economics. His research interest are in computer systems, sensor networks, computational geometry and computer gaming. He has received both an PhD and an MS in Computer Science from the NYU Polytechnic School of Engineering and his BS from Instabul Technical University.

Mema Roussopoulos is an Associate Professor of Computer Science at the University of Athens. Her interests are in the areas of distributed systems, networking, and digital preservation. She is a recipient of the NSF CAREER Award, the ERC Starting Grant Award, and Best Paper Award at ACM SOSP 2003. She received her Ph.D. in Computer Science from Stanford University and her BS from the University of Maryland at College Park.

Alex Delis is a Professor of Computer Science at the University of Athens. His research interests are in distributed computing and data management. His work has been supported by agencies and organizations in Australia, US and the European Union. He holds both a PhD and an MS in Computer Science from the University of Maryland at College Park and received his Eng. Diploma from the University of Patras.