Distributed Network Formation for n-way Broadcast Applications

GEORGIOS SMARAGDAKIS* NIKOLAOS LAOUTARIS† PIETRO MICHIARDI§
AZER BESTAVROS‡ JOHN W. BYERS‡ MEMA ROUSSOPOULOS¶

***Abstract*—In an n-way broadcast application each one of n overlay nodes wants to push its own distinct large data file to all other n-1 destinations as well as download their respective data files. BitTorrent-like swarming protocols are ideal choices for handling such massive data volume transfers. The original BitTorrent targets one-to-many broadcasts of a single file to a very large number of receivers and thus, by necessity, employs a sub-optimized overlay topology. n-way broadcast applications on the other hand, owing to their inherent complexity, are realizable only in small to medium scale networks. In this paper, we show that we can leverage this scale constraint to construct optimized overlay topologies that take into consideration the end-to-end characteristics of the network and as a consequence deliver far superior performance compared to random and myopic (greedy) approaches. We present the Max-Min and Max-Sum peer-selection policies used by individual nodes to select their neighbors. The first one strives to maximize the available bandwidth to the slowest destination, while the second maximizes the aggregate output rate. We design a swarming protocol suitable for n-way broadcast and operate it on top of overlay graphs formed by nodes that employ Max-Min or Max-Sum policies. Using measurements from a PlanetLab prototype implementation and trace-driven simulations, we demonstrate that the performance of swarming protocols on top of our constructed topologies is far superior to the performance of random and myopic overlays.**

***Index Terms*—Network formation, n-way broadcast, swarming protocols, flow networks.**

I. INTRODUCTION - OUTLINE

Motivation: The BitTorrent protocol [2] has established *swarming*, *i.e.*, parallel download of a file from multiple peers with concurrent upload to other requesting peers, as one of the most efficient methods for multicasting bulk data. A fundamental characteristic of the existing BitTorrent is that the overlay graph resulting from its bootstrap and choke/unchoke algorithms is mostly ad-hoc, in the sense that it is the outgrowth of random choices of neighboring peers. This is justified given the scale of P2P file swapping networks.

P2P file swapping, is not the "be all and end all" for swarming. In this work we consider n-way broadcasting — another class of applications, in which each one of n overlay nodes must push a very large chunk of data (a distinct file) to all other $n-1$ peers, as well as pull the $n-1$ files pushed by these other peers. Once completed, this push-pull cycle may be repeated with new sets of files.

Applications using n-way broadcasting would involve small/medium-sized networks, as they are inherently of n^2 nature. Examples include: distribution of large scientific data-sets in grid computing (*e.g.*, CERN's Large Hadron Collider data-sets), distribution of large traffic log files for network-wide distributed intrusion/anomaly detection schemes [3], synchronization of distributed databases [4], distributed batch parallel processing and storage applications [5], and several other enterprise applications. Contrary to the prevailing assumption underlying the design of BitTorrent, the nodes that make up such networks are basically cooperative (at an extreme case they belong to the same administrative authority).

Even for relatively small networks, n parallel broadcasts of distinct large files can create data volumes that are impossible to handle via centralized solutions: uploading each file to a centralized server and then copying it back to all destinations in a point-to-point manner means that the same file is transmitted $O(n)$ times over the same link, *i.e.*, imposing an $O(n)$ stress on the physical links.

n-way broadcast via swarming: Swarming is clearly an attractive approach to supporting n-way broadcast applications. The obvious solution is to outsource the push-pull functionality to BitTorrent: set-up n different torrents, each one seeded by a different node.

In this paper, we question the effectiveness of BitTorrent for n-way broadcasting (which is not what it is primarily designed to support). In particular, we note that BitTorrent runs on the topologies that result from the composition of its bootstrapping and choke/unchoke algorithms. These topologies are mostly unoptimized. Indeed, the only topological optimization in BitTorrent is a local one: under the choke/unchoke algorithm, fast peers are matched up with other fast peers *from within the same randomly bootstrapped neighborhood*. By virtue of the relatively small size of neighborhoods compared to the entire network, the resulting topology is close to being random. While randomly-bootstrapped graphs may possess desirable theoretical properties (such as small diameters), they are likely to be inefficient when compared to graphs that are systematically constructed to optimize a specific application. Notice that BitTorrent's matching of fast nodes is mostly in the protocol as an efficient tool against free-riding [6], rather than as a conscious attempt to optimize the overall overlay topology for applications such as n-way broadcast.

Our work — Optimized overlays for n-way broadcast: For n-way broadcast applications (as well as for other potential classes of applications), the main goal is to minimize the

* Deutsche Telekom Laboratories and Technical University of Berlin, Berlin, Germany.
† Telefonica Research, Barcelona, Spain.
‡ Computer Science Department, Boston University, Boston, MA, USA.
§ Networking and Security Department, Institut EURECOM, Sophia-Antipolis, France.
¶ Dept of Informatics and Telecommunications, University of Athens, Athens, Greece.

G. Smaragdakis is supported by a Strategic Research grant from Deutsche Telekom Laboratories. N. Laoutaris and P. Michiardi are supported in part by the NANODATACENTERS program (FP7-ICT-223850) of the EU. P. Michiardi is supported by the French ANR (National Research Agency) VERSO program VIPEER project. A. Bestavros and J. Byers are supported in part by a number of NSF awards, including CISE/CSR Award #0720604, ENG/EFRI Award #0735974, CISE/CNS Awards #0952145, #0524477, #0520166, and CISE/CCF Award #0820138. M. Roussopoulos is supported in part by the NoAH Specific Support Action (FP6-RIDS-011923) of the European Union. Part of this work appeared in the proceedings of the 27^{th} IEEE Infocom conference [1].

file download time by optimizing the efficiency of the entire overlay as opposed to creating a tit-for-tat environment to reign in selfish, free-riding behavior of individual nodes. Also, the scale of the applications we envision makes it possible/practical to optimize the construction of the overlay, especially if distributed optimization is used.

Armed with this realization, our goal will be to construct highly efficient topologies to be used by swarming protocols for n-way broadcast. Specifically, we construct an optimized, common overlay network, upon which swarming is used. In order to control the stress of the physical links supporting the overlay, we impose an upper bound on the degree of the nodes in the constructed overlay network.

Next we present justification for several of the salient features of our solution – features that will be developed and presented fully later in the paper.

Why swarming on top of an overlay? Because hop-by-hop relay of the entire file over a shortest-path tree embedded on the overlay topology and rooted at the seed node would take too long. We want to harness the power of parallel downloads as exemplified in BitTorrent.

Why use a common overlay? Because a topological optimization requires monitoring the performance of overlay links, and we want to amortize the cost of such monitoring — pay it only once per link and reuse the result for the benefit of all n transmissions (and avoid monitoring the same link up to n times as can happen if one builds n independent overlays).

How could swarming benefit from an end-to-end optimized overlay? Our overlays are optimized for end-to-end performance over multi-hop paths, *e.g.*, by maximizing the minimum available bandwidth to any destination over multiple paths, or by maximizing the total available bandwidth to all destinations over all available paths. From a single node's perspective, swarming involves point-to-point transfers within the neighborhood of that node. Each node, however, has in its neighborhood nodes that also belong to other "adjacent" neighborhoods. Noting this, one can see that, through swarming, data chunks eventually reach their destinations through multi-hop paths formed through single hop transfers between neighborhoods. If these multi-hop paths are end-to-end optimized, then swarming will be more effective in operating upon them as compared to upon unoptimized paths.

Why optimize the overlay based solely on network characteristics, without consideration of data availability? Arguably, one could conceive of more general overlay constructions in which neighbors are selected based on criteria involving both the network characteristics and the availability of chunks at each candidate connection point. In our work, we adopt a bandwidth-centric/data-agnostic approach to the construction of the overlay for two main reasons: (1) for large objects it is high bandwidth that leads to small delivery completion times and high object throughput; (2) the global state in terms of available chunks per node changes too frequently (with each successful chunk exchange between two nodes), resulting in an optimized topology that changes too frequently to be of practical use. The fact that we do not consider data availability in the construction of the overlay does not mean that data availability does not play a role in our approach: it does, but not at the overlay construction time-scale. Specifically, *we advocate a "two-pronged approach" operating at two distinct time scales*: at a coarse time scale, we address issues related to network characteristics through the construction of a dynamic, distributively optimized overlay, and at a finer time scale, we address issues related to data availability through the upload/download scheduling algorithms employed in the swarming protocol that runs on top of the overlay.

II. Related Work

This work is the fusion of two very recent thrusts in networking research: *network creation games* and *swarming protocols*. Network creation games appeared in computer science with the work of Fabrikant et al. [7] in which a set of nodes forms a network in a distributed manner driven by self-interest — each node pays for the creation of a number of links to other "neighbors" so as to minimize a hybrid cost that captures the purchase cost of these links and the delay for routing packets to all other destinations using own and remote links. The model targeted the creation of physical telecommunication networks through peering agreements between ISPs (hence the explicit modeling of the cost of buying a link). Laoutaris et al. [8], studied the "capacitated" version of the above problem, targeting the construction of overlay routing networks — each node is given a bound on the number of immediate peering relationships that it can establish (defined by the protocol that implements such an overlay network) and selects the best neighbors so as to minimize its sum of distances to all destinations through shortest-path routes over the resulting overlay topology. These works differ fundamentally from ours in that they target *routing*, *i.e.*, they assume that a packet from v to u is of interest only to u. Intermediate nodes w that lay on the overlay path from v to u are there just to assist in the routing of the packet. In the current work, each node is *broadcasting* a file to all destinations and thus intermediate nodes are also receivers in addition to being relay points. More fundamentally, in our case the delivery of information from v to u occurs not through a single path but (potentially) through all the available connected paths between the two end-points (because the file is cut into chunks which travel in parallel along different paths on the overlay). For this reason we employ *max-flows* as building blocks for designing the overlay (as opposed to shortest-paths which are used in point-to-point routing [7], [8]). Max-flows reflect better the nature of our application (broadcasting) as well as the nature of the employed technique for implementing it (swarming).

The BitTorrent protocol [2] has established swarming as one of the most fresh and promising ideas in contemporary networking research and thus has kick-started a (tidal) wave of research articles in the area. Our fundamental difference from this body of work, whether analytic, e.g., Qiu and Srikant [9], Massoulie and Vojnovic [10], Fernandess and Malkhi [11], Kumar and Ross [12], Sanghavi et al. [13], experimental, e.g., Bharambe [14], or measurement based, e.g., Izal et al. [15], Legout et al. [16], is that we have substituted the (close to) random graph resulting from BitTorrent's bootstrap and choke/unchoke algorithms with a highly efficient distributively optimized graph. As we show later on, such a switch boosts the performance of a swarming protocol running on top of it. We are able to obtain such highly efficient graphs because our interest is on smaller networks. We show that at such scales one can do much better than close to random.

Some other relevant works are the following ones. Massoulie et al. [17] recently showed that a simple distributed randomized algorithm can achieve the theoretical optimal broadcast rate given by Edmond's theorem [18] for a source node in a flow network. Compared to this work, we let each node select its neighbors and thus participate in the construction of the flow network, as opposed to taking it for granted. Gkantsidis and Rodriguez [19] have proposed the use of network coding as an alternative to BitTorrent's chunk scheduling algorithm. The performance benefit/added complexity ratio of employing network coding is not yet generally agreed upon [16]. Although we focus on BitTorrent-like swarming here, our optimized topologies should also benefit network-coding based swarming because they are oblivious to whether network coding is used or not. Fan et al. [20] performed a complete characterization of the different rate assignment strategies of BitTorrent-like protocols. They showed the fundamental trade-off between performance and fairness for the aforementioned protocols. Their observations provide justification for one of our design choices, which is to fully utilize the upload capacity of each node in a cooperative fashion.

Guo et al. [21] and Tien et al. [22] look at the design of multi-torrent systems. Their contribution is mostly on the measurement and the design of inter-peer incentive mechanisms for peers that participate in multiple torrents concurrently. They do not look at overlay construction issues. Interestingly, Tien et al. [22] provide justification for one of our design choices, which is to enforce that at any time there should be only one active torrent between any two nodes (more in Sect. IV). They show that deviating from this choice and allowing the transfer of multiple chunks in parallel between two nodes (one for each torrent), slows down the system by over-partitioning the upload bandwidth of nodes.

Other end-system multicast systems such as SplitStream from Castro et al. [23] and Bullet from Costic et al. [24] could be used to support n-way broadcasting by creating a separate overlay for each source. The problem with this approach is that there is no coordination across different overlays and thus there can be performance inefficiencies as well as significant overheads due to the redundant monitoring of the same physical paths multiple times from different overlays. Our approach is to construct one overlay for all sources and thus jointly optimize as well as share the monitoring cost.

The only work we are aware of on the intersection of overlay creation and BitTorrent is a very recent one from Zhang et al. [25]. It looks at the formation of Nash equilibria topologies in view of download-selfish peers that participate in a single torrent. Our overlay formation, although distributed and based on local utility functions is: (1) primarily targeting the optimization of the social utility of the network, meaning that all nodes are assumed to be under common control, and (2) considering both upload and download performance for multiple torrents, one at each node. We examine selfishness issues and how these could be addressed towards the end of our article, but this is just a supplement of our main contribution.

III. Peer-set Selection

Let $V = \{v_1, v_2, \ldots, v_n\}$ denote a set of nodes. Node v_i selects k other nodes to be in its *peer-set* $s_i = \{v_{i_1}, v_{i_2}, \ldots, v_{i_k}\}$ and establishes bidirectional links to them. Let $S = \{s_1, s_2, \ldots, s_n\}$ denote the edge set of the *overlay graph* $G = (V, S)$ resulting from the superposition of the individual peer-sets. Each link of G is annotated with a capacity c_{ij} which captures the available bandwidth [26], henceforth abbreviated availbw, on the underlying IP layer path that goes from v_i to v_j. Capacities can be asymmetric, meaning that $c_{ij} \neq c_{ji}$ in the general case. Let $MF(v_i, v_j, S)$ denote the resulting max-flow from v_i to v_j under S. Let also $\Phi(v_i, S)$ and $\Psi(v_i, S)$ denote the minimum max-flow from v_i to any other node under S, and the sum of max-flows from v_i to all other nodes under S, respectively, *i.e.*:

$$\Phi(v_i, S) = \min_{v_j \in V_{-i}} MF(v_i, v_j, S), \Psi(v_i, S) = \sum_{v_j \in V_{-i}} MF(v_i, v_j, S)$$

In the above definitions, each max-flow from v_i to an individual destination is computed independently of other max-flows from the same node to different destinations (*i.e.*, each one is computed on an empty flow network G). These definitions should not be confused with multi-commodity flow problems in which multiple distinct flows co-exist. In n-way broadcast each node is potentially a destination of a flow and the piece forwarding is difficult to analyze. Thus the classical multi-commodity flow fails to capture the dynamics of the system.

Definition 1: (Max-Min and Max-Sum peer-sets) A peer-set s_i is called Max-Min if it maximizes the minimum max-flow of node v_i, *i.e.*, $\Phi(v_i, \{s_i\} + S_{-i}) \geq \Phi(v_i, \{s_{i'}\} + S_{-i})$, $\forall s_{i'} \neq s_i$, where S_{-i} denotes the superposition of the peer-sets of all nodes but v_i. Similarly, a peer-set is called Max-Sum if $\Psi(v_i, \{s_i\} + S_{-i}) \geq \Psi(v_i, \{s_{i'}\} + S_{-i})$, $\forall s_{i'} \neq s_i$.

Lemma 1: Finding a Max-Min or Max-Sum peer-set for v_i given S_{-i} is an NP-hard problem.

Proof: See Appendices A and B. ■

These peer-set selection policies optimize the connectivity of a given node to the remaining network. One could say that this constitutes selfish behavior. This is indeed the case if the nodes use this connectivity to *only* disseminate their own file. However, when they also indiscriminately relay the files of others, which is the assumption for the applications we consider, then optimizing one's connectivity boosts the aggregate social performance of the network. Later on, in Sect. VI we study the implications when the swarming protocol (running above the overlay) ceases to be indiscriminate with respect to the upload quality it gives to local and remote files.

Why Max-Min and Max-Sum? Given a flow network G, the *broadcast problem* asks what is the maximum (broadcast) rate at which a source v_i can deliver its stream concurrently to all other nodes. Edmonds showed in [18] that the broadcast rate is equal to $\min_{v_j \in V_{-i}} \text{mincut}(v_i, v_j)$, which in view of the max-flow/min-cut theorem is equal to $\min_{v_j \in V_{-i}} MF(v_i, v_j)$. Therefore, the Max-Min peer-set is the peer-set that maximizes the broadcast rate of a node, or conversely the delivery rate to the slowest receiving peers that might be the obstacle to improve the overall performance of the system. It does so by placing the links so as to boost the max-flow to these slowest peers. Of course for this to be possible there must be available bandwidth to be utilized at the IP level (this is reflected on the c_{ij}'s which steers the peer-set selection, and which are obtained through measurements as explained in Sect IV). Edmonds gave an exponential time centralized

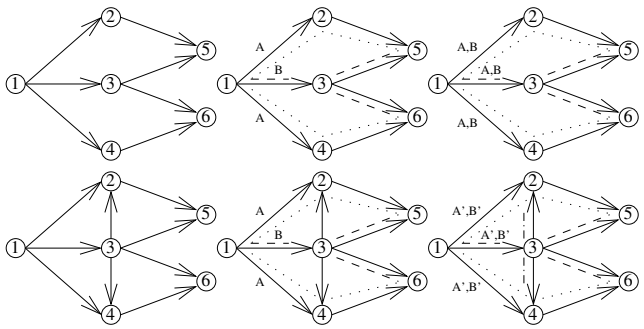

Fig. 1. Mixing max-flows. Left: empty network. Middle: RF(1,5) and RF(1,6) co-existing. Right: RF(1,2), RF(1,3), RF(1,4), RF(1,5), RF(1,6) co-existing. Top: initial network. Bottom: Initial augmented with edges, (3,2) and (3,4).

algorithm for achieving the broadcast rate, which was later improved to a small polynomial time by Lovasz, Gabow and others [27]. Recently, Massoulie et al. [17] showed that a simple randomized decentralized algorithm can achieve a delivery rate that is arbitrarily close to the broadcast rate.

A Max-Sum peer-set on the other hand is a peer-set that maximizes the *theoretical maximum aggregate transmission rate* from a node. Contrary to the Max-Min peer-set that maximizes a provably attainable broadcasting rate, the Max-Sum maximizes only an upper bound on the aggregate rate which, in the general case, is not attainable due to contention for link bandwidth when max-flows from the same source to different destinations share common overlay links.[1] We elaborate with an example.

Consider the flow network of Fig. 1 (top-left) in which all links have unit capacity and node 1 is the source. Computing each max-flow on an empty network we get that the max-flow from the source to nodes 2, 3, and 4 is equal to 1 whereas that to nodes 5 and 6 is equal to 2, thereby $\Psi(1) = 7$. Consider now the maximum real flows that can exist concurrently from the source to nodes 5 and 6 (top-center). Breaking the file into two equal parts A and B the source can transmit A at full rate over the dotted paths ($1 \rightarrow 2 \rightarrow 5$ and $1 \rightarrow 4 \rightarrow 6$) and B at full rate over the dashed path (only once over link (1,3)) and achieve concurrent real flows that match the capacity of corresponding max-flows on an empty graph, *i.e.*, RF(1,5)=MF(1,5)=2 and RF(1,6)=MF(1,6)=2. This is possible because a single transmission of B on the edge (1,3) suffices for contributing to both RF(1,5) and RF(1,6). Thus the two flows don't compete for bandwidth on the shared link and can achieve the same capacity as the corresponding max-flows on empty networks. This is not, however, generally possible. On the top-right part of the figure we depict the situation when sending from the source to all destinations (nodes 2-6) concurrently. In this case the entire file (both A and B) has to go over links (1,2), (1,3), and (1,4) and thus RF(1,5)=RF(1,6)=1<MF(1,5)=MF(1,6)=2 leading to a real aggregate rate $\tilde{\Psi}(1) = 5$ smaller than the bound $\Psi(1) = 7$.

Generally, the bound becomes less tight with increasing link density k/n. On the bottom-left part of Fig. 1 we add to the previous network two new links: (3,2) and (3,4). It is easy to verify that the max-flow from the source to nodes 2, 4, 5, and 6 is now 2 and to node 3 is 1, leading to $\Phi(1) = 9$. As before, if we consider only the flows to 5 and 6, it is easy to see that their max-flow values can co-exist. Considering, however, the flows to all destinations, we see that any partition of the file into parts will inevitably lead again to all real flows being 1, whereas the corresponding max-flows with the exception of MF(1,3) are now 2.[2] In other words, although the new links increased both MF(1,2) and MF(1,4) by 1 compared to the previous network, they cannot increase any of the real flows and thus widen the gap between the bound ($\Phi(1) = 9$) and the maximum attainable aggregate rate ($\tilde{\Phi}(1) = 5$).

To sum up, we propose and study these peer selection policies for the following reasons: (1) Max-flows are used to capture the fact that in a swarming protocol the chunks of a source node v_i travel towards a sink node v_j over (potentially) all the available paths of the overlay graph of point-to-point peer relationships. (2) The gap between the bound on the aggregate rate $\Psi(v_i, S)$ given by a Max-Sum peer-set and the actual maximum attainable aggregate rate $\tilde{\Psi}(v_i, S)$ which factors in the sharing of overlay links from multiple max-flows to different destinations, is reduced by the fact that swarming protocols guarantee that any chunk is transmitted at most once between any two peers; therefore, $\tilde{\Psi}(v_i, S)$ can use an overlay link multiple times (for different max-flows) but would seize bandwidth only once, thereby reducing its gap from the bound $\Psi(v_i, S)$ that assumes that the entire flow network is available to each individual max-flow from v_i. (3) The overlay network has to be rather sparse (small k) so as to limit the stress on the physical links. Thus the bound Max-Sum won't be very much off from the actual achievable aggregate rate and it makes sense optimizing the peer-set based on it. Regarding Max-Min, this is provably attainable, and optimal for broadcast rate as discussed earlier.

Since a node cares to both upload its local file to all other nodes as well as download from them all remote files, we combine the previous definitions in the following objective functions:

$$\hat{\Phi}(v_i, s_i) = \alpha\Phi(v_i, \{s_i\} + S_{-i}) + (1-\alpha)\min_{v_j \in V_{-i}} MF(v_j, v_i, \{s_i\} + S_{-i}),$$

$$\hat{\Psi}(v_i, s_i) = \alpha\Psi(v_i, \{s_i\} + S_{-i}) + (1-\alpha)\sum_{v_j \in V_{-i}} MF(v_j, v_i, \{s_i\} + S_{-i})$$

In the above functions, the parameter α regulates the relative importance between upload and download quality in selecting a peer-set. If the link capacities are symmetric, then optimizing $\hat{\Phi}$ or $\hat{\Psi}$ reduces to optimizing Φ or Ψ, independently of α.

IV. Node Architecture

Nodes consist of the following components: a *peer selection module* implementing the peer-set selection algorithms described in Sect. III; a *downloader module*, responsible for issuing requests to neighboring nodes and downloading missing chunks; and an *uploader module*, responsible for sending back local and in-transit chunks (an in-transit chunk is a chunk that does not belong to the local source file). In this section we describe these three modules under the assumption that nodes are cooperative (therefore we don't need mechanisms like choke/unchoke). Later on, in Sect. VI we relax this assumption.

[1] The contention between max-flows "from" different sources does not come explicitly in these objective functions. It is captured in our framework through the measured availbw c_{ij}: the availbw on a direct overlay link from v_i to v_j depends on the capacity of the underlying physical path *and the amount of this capacity already captured by the competing max-flows from other sources. At this level the problem is indeed a multi-commodity flow.*

[2] The fact that the entire file has to go over the edge (1,3), eliminates any chance for increasing the real flows to nodes 2, 4, 5, and 6 beyond 1.

A. Peer Selection Module

Every time period T, a node: (1) measures its available bandwidth to all other nodes using `pathChirp` [28], (2) executes a peer-set selection algorithm from Sect. III and connects to the corresponding nodes (incoming links are left untouched). Since both Max-Min and Max-Sum are NP-hard, we use a fast local-search heuristic [29] to compute approximately optimal peer-sets. Each node initially selects k other nodes as peers. Each node is allowed to periodically drop the connection to a uniformly at random selected peer and establish a connection with a non-peer node to improve its connectivity. In our implementation of local-search each node connects to the best candidate of non-peer nodes. We verified to be always within 1% of the exact optimal for all problem sizes on which we were able to use integer linear programming to compute the latter. Once links are established, the node keeps monitoring them (including the incoming ones) and relays their capacity to all other nodes through an overlay link-state announcement protocol. Remote nodes need this information to compute their own peer-sets. Although each node measures $O(n)$ overlay links every re-wiring epoch T, the monitoring and announcement overhead is only $O(kn)$ and not $O(n^2)$ since only the $O(k)$ established links are monitored and announced in between the (infrequent) rewiring epochs, where $k \ll n$.

B. The Downloader Module

The downloader module monitors the available chunks on the peer-set and issues requests for downloading missing ones. The selection is based on the well established *Local Rarest First* (LRF) heuristic [16] that looks at the peer-set and issues a request for any missing chunk that is among the least replicated ones in the peer-set. New requests are triggered either upon the completion of a download, or if an overlay link is inactive, upon the detection on the other side of the link of a missing chunk.

C. The Uploader Module

The uploader receives requests and sends back chunks. Our baseline uploader allows for *up to 1 active upload (chunk) per overlay link (neighbor)*. It implements this by maintaining a FIFO queue for each overlay connection. This choice bounds the number of concurrent uploads by the number of neighbors thereby avoiding excessive fragmentation (over partitioning) of the upload bandwidth of the local (physical) access link of a node (this choice is backed-up by results appearing in [22]). We also experimented with an uploader that allows up to 1 active chunk per source file per connection, but this can lead to up $n-1$ parallel uploads per overlay link, which becomes problematic as n increases. Indeed, over-partitioning the upload bandwidth defeats the entire concept of swarming: it takes too much time to upload an entire chunk, and during this time the downloading node is under-utilizing its upload bandwidth as it cannot relay the chunk before it completes the reception. We want to note, however, that our baseline design is by no means claimed to be optimal. For an example, consider a node that can upload to its first $k-1$ neighbors with rate x and to the last one with rate larger than $k \cdot x$. Then as long as this last neighbor can always find k missing chunks from our node, and can also itself disseminate them further down in the network faster than the $k-1$ slow neighbors, then the system would be better off allowing up to k parallel uploads to the fast one at the expense of the slow ones. Such situations though are rather peculiar and even if they arise, it is difficult to check the necessary conditions for taking advantage of them, so we leave their investigation to future work and stick to the simple one-chunk-per-connection policy.

V. Performance Evaluation

In this section we compare the performance of Max-Min and Max-Sum peer selection policies against three reference selection policies: Random (node v_i selects k peers at random from the set of all nodes in V_{-i}); k-Widest (node v_i selects node v_j if c_{ij} is among the k largest ones across all nodes in V_{-i}); Rand k-Widest (v_i performs k-Widest on a random subset of V_{-i} of size $\beta \cdot k$). Rand k-Widest is included in the evaluation to mimic the effect of combining random bootstrapping with choke/unchoke in BitTorrent.[3]

We compare these policies in terms of (node,remote file) *finish times*. We denote $f(j,i)$ the time that the sink v_j completes downloading the file of source v_i, assuming that all exchanges start at time 0. In all experiments we assume that nodes are fully cooperative (they belong to the same authority) and thus follow exactly and truthfully the peer-selection policies of Sect. III and the swarming protocol of Sect. IV (*i.e.*, no choke/unchoke mechanism is employed). We discuss the impact of selfishly behaving nodes in Sect. VI.

Our performance evaluation is done in two settings. In the first, we assume that the n-way broadcast is to be carried over the Internet. We do so by evaluating the performance of a prototype implementation of our architecture on PlanetLab. In the second, we assume that the n-way broadcast is to be carried on a closed (controlled/isolated) network. We do so by evaluating the performance of a prototype implementation of our architecture on a discrete event simulator of the closed network.

A. Case Study 1: A PlanetLab Prototype

In this setting, we compare the performance of different overlay topologies when the underlying physical network is the Internet and the overlay nodes are single-homed, *i.e.*, all overlay links of a node go over the same physical access link. For this purpose we selected $n = 15$ PlanetLab nodes. The distribution of nodes is as follows (we tried to use operationally stable and geographically diverse node set) : ten in North America[4] , one in South America[5] , three in Europe[6] , and one in Asia[7] . Each one of the aforementioned nodes disseminated a unique 100MByte file and was allowed to connect to $k = 2$ neighbors (and accept additional incoming links). Notice that we limited our experiment to only 15 nodes and only 100MBytes per node so as to keep the amount of exchanged traffic on PlanetLab at reasonable levels, while also allowing us to monitor the

[3] Unless otherwise noted, we used $\beta = 2$.

[4] planetlab4.csail.mit.edu, planetlab2.millennium.berkeley.edu, planetlab2.utep.edu, planetlab2.acis.ufl.edu, planetlab-8.cs.princeton.edu, planetlab-2.cs.colostate.edu, planetlab5.cs.duke.edu, planetlab1.cs.northwestern.edu, planetlab3.flux.utah.edu, planetlab01.cs.washington.edu

[5] planetlab-02.ece.uprm.edu

[6] planet2.zib.de, planet2.colbud.hu, planetlab3.xeno.cl.cam.ac.uk

[7] planetlab1.netmedia.gist.ac.kr

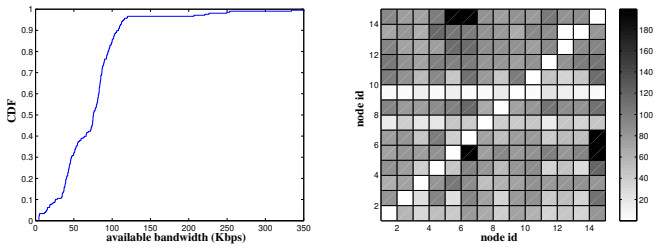

Fig. 2. PlanetLab experiment, empirical CDF and scatter plots of available bandwidths.

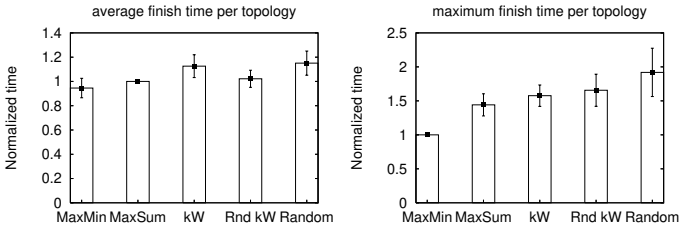

Fig. 3. PlanetLab experiment, performance evaluation of different wiring strategies.

network throughout the experiment. Notice that if data were to be transferred in a point-to-point manner, then it would amount to over a Terabyte for each execution of the entire experiment: 5 different peer-set selection policies, each one generating 15·14·100MBytes of data at each run, and repeated 10 times to get confidence intervals (the experiment was performed between June 4th and June 30th, 2007). We let the re-wiring epoch be $T = 10$ minutes and the measurement/announcement epoch for existing links be 2 minutes. Also we set $\alpha = 0.5$ to indicate that nodes care equally for download and upload quality. In all our experiments we used `pathChirp` [28], an end-to-end active probing tool which requires the installation of sender and receiver modules of the aforementioned tool in each node. `pathChirp` is a lightweight, fast and accurate tool, which fits well with the PlanetLab-specific constraints; namely it does not impose a high load on PlanetLab nodes since it does not require the transmission of long sequences of packet trains and does not exceed the max-burst limits of PlanetLab. The additional overhead of the tool in terms of bandwidth consumption is negligible and does not affect the performance of the content distribution. We limited the maximum experiment duration to 10 seconds per peer (thus a full estimation for the available bandwidth from any node to all the other nodes was achieved in approximately 2 minutes) and we used as available bandwidth the average available bandwidth (per peer) observed during the experiment. In Fig. 2 we plot the cumulative distribution function (CDF) of the pairwise available bandwidth as well as the scatter plot illustrating the available bandwidth among nodes of a typical experiment in PlanetLab. The diversity of available bandwidth between peers that observed was moderate. We did not observe huge variability of the available bandwidth while performing our experiments (variability was limited to the available bandwidth among a few nodes only).

To perform the experiment, we modified both a BitTorrent client and tracker. We used the `mainline 4.0.2` BitTorrent client. We disabled the choke, unchoke and optimistic unchoke functionality and we set no limits for both the upload and download rate as well as the number of active peers. Although we are aware of the intrinsic limitations of PlanetLab as well as the PlanetLab policy of fair sharing of bandwidth among slices that use the same node, we were able to achieve very high upload and download rates (close to the estimated available bandwidth). To minimize the interaction of our experiment with other bandwidth demanding experiments, we performed the experiments after monitoring the activity of competing slices for bandwidth in the selected nodes.

We used the `phpbttrkplus-2.2` BitTorrent tracker, which is a php-based tracker that maintains records about the activity of nodes (in a `mysql` database). We installed the aforementioned tracker in one of our machines (egoist.bu.edu), and we modified it to reply to requests initiated by nodes, by providing the summary of the requested peer set (ip, port, status) and not of a random peer set (as was initially designed).

For a node v_j, we compute its maximum finish time $\max(j) = \max_{i \neq j} f(j,i)$, *i.e.*, the time at which it has completed downloading *all* $n-1$ remote files, as well as its average finish time $\text{avg}(j) = 1/(n-1)\sum_{i \neq j} f(j,i)$. For peer-set selection policy $\mathcal{X}$, we let $\max(\mathcal{X}) = \max_j \max(j)$ denote its *maximum finish time* across all nodes, and $\text{avg}(\mathcal{X}) = 1/n \sum_j \text{avg}(j)$ denote its *average finish time* across all nodes.

On the left-hand-side of Fig. 3 we present the normalized average finish time of each policy with respect to the average finish time of the Max-Sum policy. On the right-hand-side, we present the normalized maximum finish time of each policy with respect to the maximum finish time of the Max-Min policy. These results show that the various policies perform quite similarly with respect to average finish time. When looking at maximum finish times though, the picture is completely different. Max-Min manages to complete all downloads anywhere between 40% and 120% faster than the heuristics and almost 30% faster than Max-Sum. This can be very significant for Bulk Synchronous Parallel (BSP) applications [30], in which the global progress depends on the finish time of the slowest node. It is worth noticing that optimizing the maximum finish time is much more difficult than optimizing the average, and thus it should come as no surprise that the heuristics perform well on average but fail to improve the worst case.

B. Case Study 2: A Dedicated Network Prototype

In this setting, we examine overlay networks whose links are dedicated, meaning that they do not compete for bandwidth in the underlying physical network. This model is plausible for (multi-homed) networks set up in support of an enterprise through the acquisition of dedicated links to connect its various locations. Such link acquisitions could be done through SLA contracts with ISPs or through virtualization technologies such as those envisioned for Cabo [31] or GENI.[8] In either cases, a dedicated link could be set up between two enterprise nodes i and j for a given price. Any such dedicated link will have a nominal capacity c_{ij}, which may depend on any number of factors (*e.g.*, physical constraints of the underlying technology, the demand at the ISP for carrying traffic between these two locations, or the price paid for various links). Since setting

[8] GENI: Global Environment for Network Innovations. `http://www.geni.net`

up a complete network to connect all n nodes directly to each other may not be feasible (especially for systems of moderate sizes), designers of such enterprise networks are likely to construct the network so as to maximize its utility with respect to some objective function. Independent of which process/strategy is used to construct the optimized overlay, the resulting network would allow all enterprise nodes to communicate either directly or through overlay paths.

The construction we propose for optimizing the overlay for n-way broadcast proceeds as follows. First, we order the nodes according to their ids (ids are assigned, in increasing order, by selecting the overlay nodes uniformly at random). Next, we proceed in rounds in which nodes take turns in selecting their peer-sets (as discussed in Sect. III and IV). This process is repeated until we converge by reaching a round that does not introduce changes in the constructed overlay topology.[9]

Towards our goal of evaluating the impact of various peer selection policies on the performance of n-way broadcast in this setting, we developed a discrete-event simulator that is able to run over dedicated overlay networks. We constructed the dedicated overlay (enterprise) network using the procedure described above, using the publicly available trace of Sprint physical topology taken from Rocketfuel [32].[10] In particular, we assumed that the dedicated capacity that could be acquired from the ISP (Sprint) would reflect an "equal-share" partitioning, which we approximated as follows. We counted the number of shortest-paths (for all physical node pairs) that go over a physical link and set the available bandwidth of that link to be its real capacity divided by this number.[11] Then, for an overlay link (i, j) we set $c_{i,j}$ to be equal to the available bandwidth of the tightest physical link on the induced shortest-path over the physical topology. This produces the amount of available bandwidth that the ISP can guarantee for the new application if it admits it into its network and treats it equally with pre-existing ones. In Fig. 4 we plot the CDF of the pairwise available bandwidth as well as the scatter plot illustrating the available bandwidth among nodes of a typical experiment. The diversity of available bandwidth between peers is more intense (compared to the PlanetLab experiment), as there are nodes which are connected to other nodes achieving very high available bandwidth and others that are connecting achieving very low available bandwidth.

One advantage of simulations (compared to PlanetLab prototyping) is that it allows us to consider a bigger network. In particular, in the experiments that follow, we study overlays of size $n = 50$ nodes, which are selected uniformly at random[12]

[9] It is worth mentioning that the convergence of the above procedure relates to a question regarding the existence of pure Nash equilibria, and their reachability through local improvement paths, in a strategic game with Max-Min or Max-Sum as its payoff function. Although interesting from a theoretical standpoint, the question is not directly relevant here as we have assumed that nodes forward indiscriminately local and in-transit chunks. In all our experiments we got fast convergence but could also stop prematurely after a maximum number of iterations so as to deal with inexistence, unreachability, or slow convergence to stable overlay topologies.

[10] The topology was inferred using the methodology described in [32]. The link weights we used for the shortest path algorithm are those inferred in [33]. The capacities of the links were made publicly available by Sprint.

[11] The idea is that each pair of physical nodes represents a different application that is assigned an equal share of the physical capacity of all links on which it competes with other applications.

[12] The CDF of available bandwidths for the sampled set is similar to that of the entire set of nodes in the Sprint physical topology.

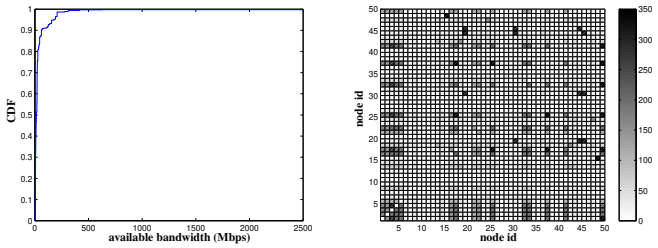

Fig. 4. Sprint physical topology, empirical CDF and scatter plots of available bandwidths.

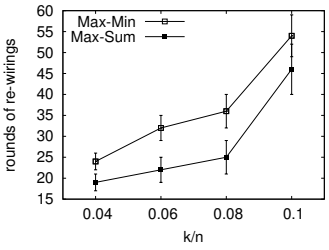

Fig. 5. Sprint physical topology, speed of convergence to a stable overlay topology with 50 overlay nodes using different peer-selection policies.

from the physical SPRINT network — each node holding a 500MByte file. As in the PlanetLab prototype, there is no notion of choke, unchoke and optimistic unchoke. The local piece selection follows a rarest first policy, there is no limit in the upload and download rate and the files are cut into 256KByte chunks (that maintains blocks of 16KBytes which is the actual transmission unit).

In Fig. 5 we plot the speed of convergence from an arbitrary initial overlay topology (100 topologies generated) to a stable overlay topology for different link densities (k/n). Fig. 5 shows that a stable overlay topology can be found within a small number of iterations, especially when the link density is small. The speed of convergence can be further improved if we allow re-wirings only when the improvement in performance exceeds a particular threshold. Moreover, the figure shows that there is a diminishing return in performance gain due to re-wiring in high link density overlays after a small number of iterations.

In Fig. 6 we compare the average and maximum finish times of different policies for different link densities (k/n). Compared to the previous results from PlanetLab, we observe a qualitatively similar behavior. The gap, however, between Max-Min and the rest in terms of maximum finish time widens substantially: Max-Min is able to finish 2-3 times faster in this setting, even for relatively large k/n ($\sim$10%). The reason is that Max-Min has more real bandwidth to work with in this case. When it places a link (i, j), the capacity (both upload and download) of the two endpoints increases by the capacity of the newly-added dedicated overlay link, whereas in PlanetLab the physical bandwidth is fixed, so when Max-Min places an overlay link it can only benefit by whatever unused bandwidth exists in the underlying physical network.

It is worth mentioning that Max-Sum may lead to poor performance when the ratio k/n is low.[13] This is expected

[13] This should not be confused with the discussion in Section III on the tight bound of Max-Sum under low link density.

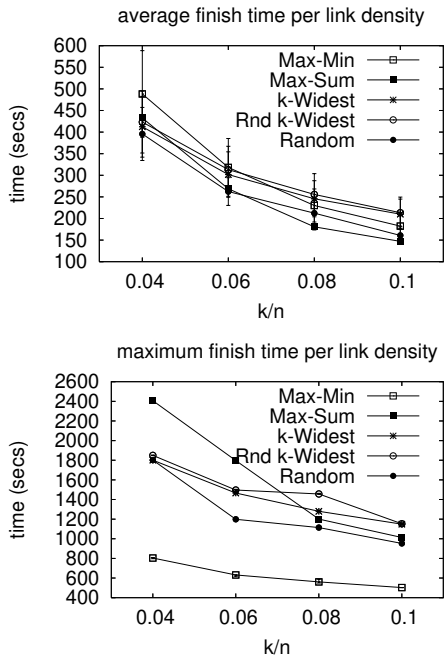

Fig. 6. Simulation of a closed overlay network based on Sprint physical topology.

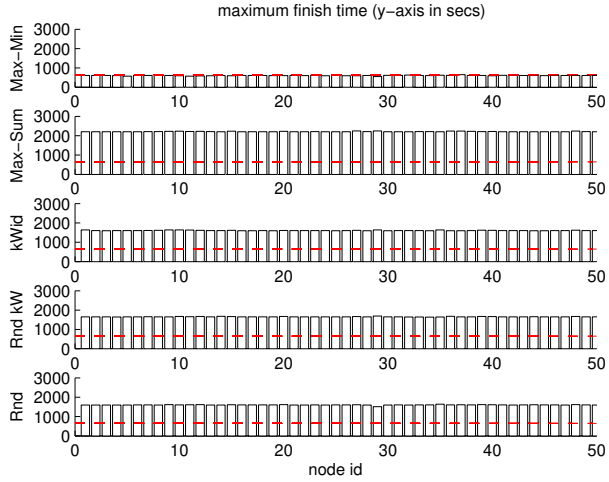

Fig. 7. Maximum finish time per node based on Sprint physical topology with link density $k/n = 0.04$. The horizontal dashed line indicates the maximum finish time on the Max-Min topology.

as the rational behind the Max-Sum wiring strategy is to maximize the average maximum flow from one node to all the other nodes. Nodes that do not contribute significantly to increasing the maximum flow are not popular, thus few connections are established (by other overlay nodes) to these nodes. As more network resources (links) are allowed to be available to overlay nodes, they establish connections that do not contribute a lot to the maximum flow, improving the maximum finish time. This is observed in Fig. 6 (bottom); the maximum finish time of Max-Sum decreases significantly as the link density increases. Similarly observations are observed for the average finish time per file per node (see Fig. 6 (top)), although there are no significant differences among the performance of different wirings.

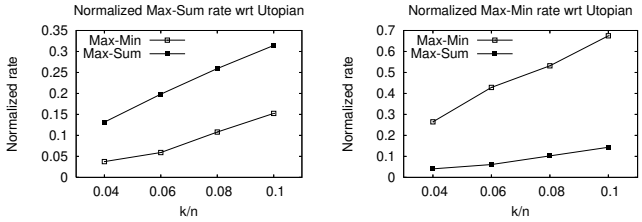

Fig. 8. Normalized Max-Sum and Max-Min rate with respect to the Utopian one based on Sprint physical topology.

Another important observation is that under any wiring strategy the maximum finish time of the nodes is almost identical (see Fig. 6(bottom)). This is another indication that the finish time is dominated by the slowest pieces (see also Fig. 7). Moreover, the performance of k-Widest may be worse than the performance of Rand k-Widest, as a globally greedy selection of peers may penalize the slowest peers more than a local greedy selection.

To characterize the graphs obtained by Max-Min and Max-Sum, we compare them with an optimal centralized construction (which may not be feasible). Let $maxout(v)$ be the sum of the bandwidths of node v's outgoing links (assuming that node v established k links and $n-k-1$ links are established by other nodes. In a social optimal graph, node v's total output rate cannot exceed

$$s(v) = \sum_{\forall u \neq v} min(maxout(v), maxout(u))$$

Define the $\sum_{\forall v} s(v)$ as the *Utopian* Max-Sum social rate. Define as the *Utopian* Max-Min rate the value of

$$s^{min} = \min_{\forall u,v} (min(maxout(v), maxout(u)))$$

The Max-Sum and Max-Min social rates are defined accordingly for any wiring where m_i $(n-1 \geq m_i \geq k)$ links are used by any node v_i on a given topology. In Fig. 8, we illustrate the Max-Sum and Max-Min social rate obtained by the Max-Sum and the Max-Min wiring normalized by the Utopian Max-Sum and Utopian Max-Min social rate respectively (left and right figure respectively), for different values of link density. Both the Utopian social rates increase with link density and Max-Min social rate of the Max-Min wiring is close to the Utopian social rate even for low link density.

In Fig. 9, we illustrate the node degree; on the x-axis nodes are ranked according to each $maxout$, i.e. the node with the lowest $maxout$ is ranked last for low link density (qualitatively similar observations are obtained for higher link density). In the Max-Min topology, nodes with low $maxout$ have high degree. Note that simple heuristics like link establishment between any node with the node with the lowest $maxout$ may be useful only in the extreme scenario where there is only one node with low $maxout$ (as we will comment in the next section). In general, we observe that the distribution of the degree of the nodes is related to the distribution of the $maxout$, and thus it is difficult to construct heuristics that can work well in practice. We found that the above observation holds for topologies with higher link density.

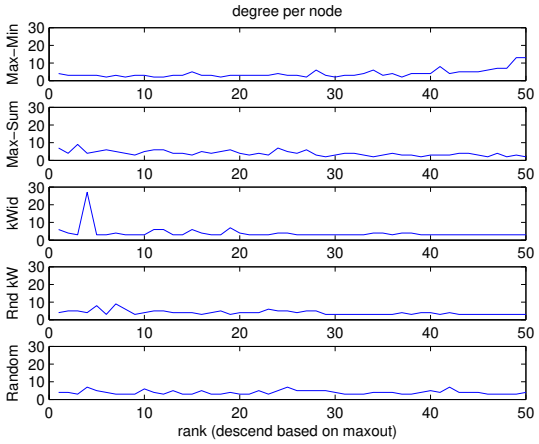

Fig. 9. Node degree on different overlay topologies for $k/n = 0.04$.

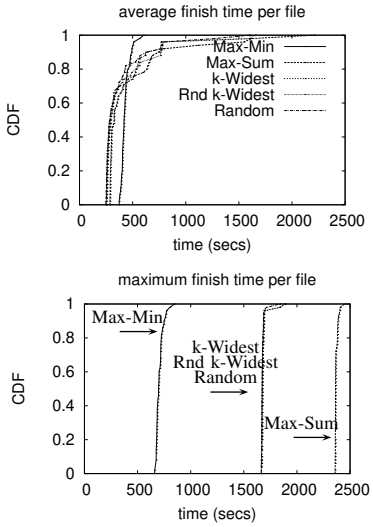

Fig. 10. CDF for the average and maximum finish time per file per node for link density $k/n = 0.04$.

Turning our attention to the average and maximum time for a document to be disseminated, we observe that the Max-Min wiring strategy has the tendency to (slightly) increase the average time that a file needs to be disseminated, but the decrease of the maximum time of any file to be disseminated is significant. The delay is mainly due to the injection of rare pieces by the slowest node or nodes. To depict this, we plot the CDF of the average and maximum finish time for low link density (see Fig. 10). Qualitatively similar observations are obtained for higher link density. Finally, an important observation is the continuous availability of requested pieces (thus the assumption of utilized parallel downloads with TCP is valid). This is consistent with observations we made in the PlanetLab prototype. This is expected, as in contrast with the case of a single torrent, pieces from different files are distributed among the nodes.

C. Case Study 2b: A Dedicated Network Prototype with a very slow node

We study the case where there is a very slow node (the $maxout$ of this node significantly deviates from the value of

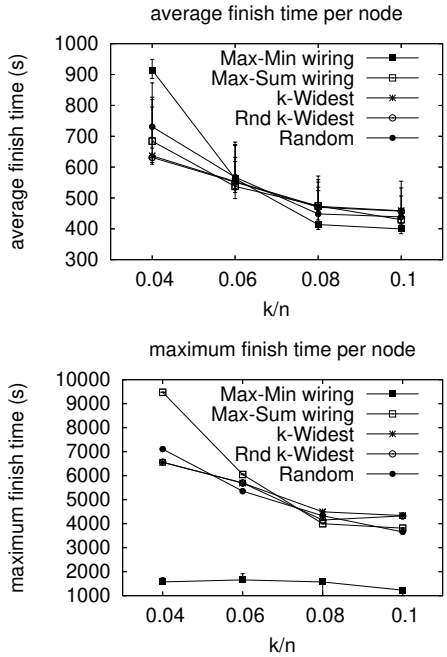

Fig. 11. Simulation of a closed overlay network based on Sprint physical topology, in the presence of a very slow node.

$maxout$ of the other nodes) based again on the Sprint physical topology. In Fig. 11, we illustrate the average and maximum finish time under different wiring strategies. In the presence of a very slow node, the performance of Max-Min topology is superior compared with the performance of the other wiring strategies, for average and especially for the maximum finish time for high link density. In Max-Min topologies, nodes are able to download all the file 3-6 times faster than in any other topology, even when the link density is high. Moreover, as we show in Fig. 12, the average delay that is introduced for the dissemination of the documents, except the one that is uploaded by the slowest node, is negligible. On the other hand the improvement of the maximum finish time in the Max-Min topology is significant. In the presence of a very slow node the performance of the Max-Sum can be very bad. In the aforementioned setting we observed that the performance of a simple heuristic where each node establishes connections with the slowest node may improve the maximum finish time. However, Max-Min provides the lower maximum finish time as it takes into consideration the capability of each node while keeping the degree distribution close to uniform.

VI. Dealing with Selfish Behavior

Up to now we have assumed that nodes are fully cooperative, which is a realistic assumption for the applications enumerated in the Introduction. In this section we explore ways to accommodate applications that involve selfish nodes. We focus on the following definition of selfishness:

Definition 2: (Upload-selfishness) An *upload-selfish* node is a node that wants to use as much of its upload capacity as possible for forwarding its local chunks and avoid "wasting" it on relaying any in-transit chunks that it holds.

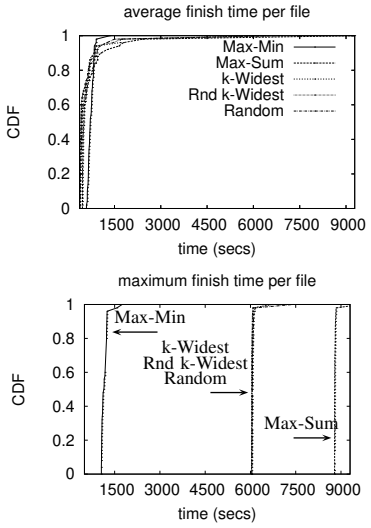

Fig. 12. CDF for the average and maximum finish time per file per node for link density $k/n = 0.04$, in the presence of a very slow node.

A. A Brief Taxonomy of Deterrence Mechanisms

The amount of extra benefit for an upload-selfish node (and potential harm to others) depends on the mechanisms that the network deploys for discouraging such behavior. We examine the following cases.

Case 0 (neutral): Here the network stays neutral and does not deploy any deterrence mechanism. In such a setting, the upload-selfish node could simply upload its own chunks and ignore all other requests. The harm to cooperative nodes can easily be quantified for this case, so we do not discuss it further; it will be proportional to the number of upload-selfish nodes. Cooperative nodes will be slowed down and at an extreme case will be unable to receive some files (*e.g.*, when all their neighbors are upload-selfish, which is similar to the case of an Eclipse attack [34]).

Case 1 (oblivious retribution): A network can employ several retribution mechanisms to punish a node that fails to deliver a chunk after a request. The choke/unchoke [2] mechanism of BitTorrent or modified versions based on bit-level tit-for-tat [14], [21] are two established existing proposals. Contrary to the original BitTorrent, such mechanisms are marginally useful here because they are oblivious to whether a node uploads local or in-transit chunks. An upload-selfish node will appear to be contributing by the mere fact that it is uploading its own chunks. Thus oblivious strategies fail to punish nodes that "free-ride" by not uploading in-transit chunks.

Case 2 (non-oblivious retribution): Assume that there exists a non-oblivious retribution mechanism that punishes a node that fails to service requests[14] for in-transit chunks that it holds. What can a selfish node do against such a mechanism? The simplest strategy is to hide (by not announcing) the availability of in-transit chunks it holds and thus get rid of the burden of having to service requests for these chunks. This can be addressed with a simple *two-hop announcement strategy* in which a node that uploads to another node announces on its behalf the availability of the chunk (using HAVE messages [2]) to downloaders belonging to the peer-set of the receiving node. This requires obtaining upon bootstrap (and re-wiring) second hop neighbors. Assuming that the retribution is severe enough, the upload-selfish node will have to honor all requests. Nonetheless, the upload-selfish node still has some room to game the system by changing the uploader and the downloader as follows.

– The upload-selfish node can substitute each FIFO queue at its uploader with a *selfish FIFO* (S-FIFO) that gives priority (preemptive or non preemptive) to requests for local chunks.

– The upload-selfish node can switch from Least Replicated First to Most Replicated First (MRF) downloads. Highly replicated chunks receive fewer requests and thus reduce the "waste" of upload bandwidth for sending in-transit chunks is smaller since most nodes already have these chunks and any requests for these chunks will be divided over many peers.

Since it is difficult to detect such deviations from the protocol, we instead quantify their impact.

B. Quantifying the Impact of S-FIFO/MRF

We quantify the advantage for a single upload-selfish node by looking at the ratio between the time it takes to upload its file to all other nodes when it is selfish and when it is cooperative, granted that all other nodes are cooperative. We examined this ratio for different overlays built on the Sprint trace and for different choices with respect to the choice of selfish node. We consider three cases, where the selfish node is : (1) the *slowest* node, *i.e.*, the one whose adjacent links have the minimum aggregate upload capacity; (2) the *fastest* node; or (3) a *typical* node (median upload capacity).

On the Max-Min overlay the selfish node reduced its maximum upload finish time by 30% when it was the slowest one. There is also, on average, a 15% reduction on the maximum finish time of all the other nodes. When it was a typical (or the fastest one), then it got almost no benefit, since in these cases the bottleneck is at the downloading nodes (so local selfish behavior cannot help). In all other overlays, the selfish node got almost no benefit, even when it was the slowest node. Unlike the Max-Min, the other overlays are not optimized for the slowest node, so even if this bottleneck node tries to selfishly upload its file, it cannot really benefit because it has very limited bandwidth.

From the above, it is clear that there exist cases in which upload-selfishness pays substantially. Given that upload-selfishness is hard to detect, we now examine its impact on cooperative nodes. We consider again a single selfish node (one can easily extrapolate for multiple selfish nodes). The impact depends on the considered metric and on the identity of the selfish node. If we care about the worst-case finish time of cooperative nodes and let the selfish node be the slowest node, then counter-intuitively, the impact on the cooperative nodes is positive. This is because by being selfish, the slowest node helps all other nodes improve their (bottleneck) downloads from it. To depict this we show a scatter-plot on the first row (left plot) of Fig. 13 with the finish time for each pair (node,remote file) when the topology is random and all the nodes are cooperative. The solid black line that stands out corresponds to the slowest node (node 29), whose file is the

[14] We do not want to punish nodes that do not have enough in-transit content for whatever reason (slow local link or peer-set) but would relay if they had, so we only punish when a request exists and is not honored.

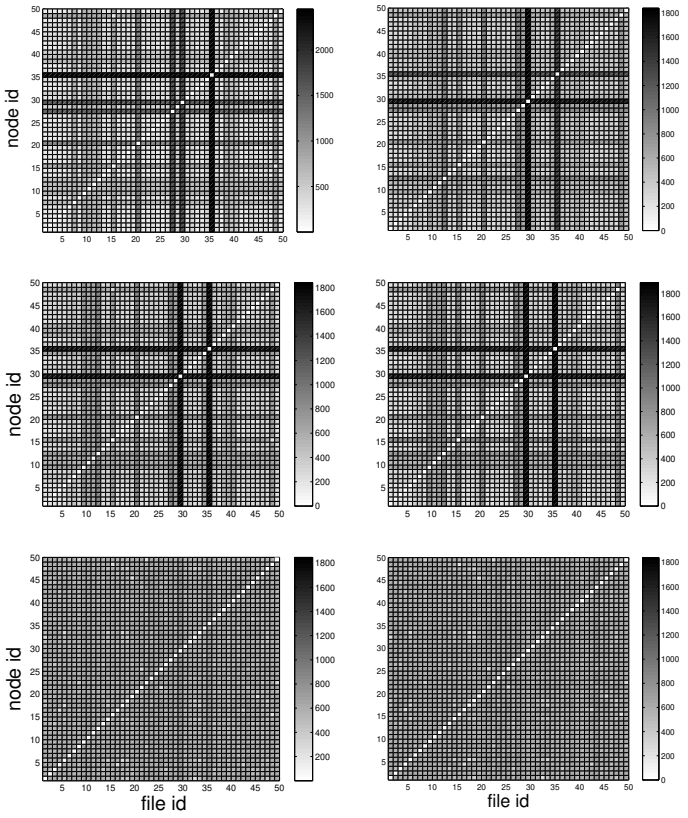

Fig. 13. Maximum finish time for all nodes and all files under different wirings (from left to right): MaxSum, Random, k-Widest, Rnd k-Widest, MaxMin strategies with cooperative nodes and Max-Min with upload-selfish slowest node ($k/n = 0.04$).

last one to be downloaded by all others. Qualitatively similar observations are obtained for all other wiring strategies[15] except Max-Min (see the first two rows of Fig. 13). To contrast this, we plot in the last row the corresponding times when the topology is Max-Min and all nodes are cooperative (bottom left plot) and when the topology is Max-Min and the slowest node is upload-selfish (bottom right plot). A first observation is that the Max-Min topology does a pretty good job at smoothing out the differences in the maximum finish times with slight increase on the average finish time (note that some cells may be darker on the Max-Min topology compared to the corresponding cells on other topologies). As it can be seen, the combination of Max-Min topology and upload-selfish scheduling on the slowest node (see last row, right plot) does even a better job at smoothing out the differences in maximum finish times. If, on the other hand, the selfish node is a typical node or the fastest node, then its effect on the download quality of others is marginal. First, its own file is not a bottleneck. Second, the relay of in-transit chunks is largely carried by the other $n-1$ nodes. Third, S-FIFO and MRF impact primarily first-hop neighbors and have small impact on nodes further away. Although we present the scatter-plot for low link density ($k/n = 0.04$) qualitative similar observations are observed for higher link densities. Qualitatively similar observations are obtained even in the presence of a very slow node (node 44), as it is illustrated in Fig. 14.

[15] Note that the slowest node may not be the same among different topologies.

Overall, upload-selfishness, unlike its name suggest, is not necessarily bad. A socially inclined global scheduling policy, for example, would certainly make slow nodes upload only their own chunks so as to reduce the severity of the bottlenecks that they cause. More generally, for social optimality, one should split the upload bandwidth of a node between local and in-transit chunks according to the relative speed of the node. Nodes who are fast should contribute heavily in relaying in-transit chunks. Nodes who are slow, should focus only on uploading their own chunks so as to avoid becoming bottleneck points. Stated differently, *a single uploading policy across all nodes cannot be socially optimal*.

C. Download-Selfishness

It is tempting to ask whether a notion of *download-selfishness* would make sense. Our answer leans towards the negative. First, there is no contention between local and in-transit chunks in the incoming direction towards a node — only in-transit chunks flow there. Second, as long as the downloader keeps all its overlay connections busy by immediately identifying and requesting missing chunks, its download-finish time will be the same, so it gets no foreseeable benefit by deviating from LRF. Finally, trying to manipulate the system by advertising false c_{ij}'s for the established links can be disclosed by having nodes periodically "audit" others by measuring some remote c_{ij}'s and comparing with the advertised values on the link-state protocol. Such methods are quite elaborate and fall outside the scope of this paper.

VII. The Effect of Delayed File Request and Availability

Up to now we have assumed that the file download requests are perfectly synchronized and all the files are immediately available. This might not be always the case either due to dissimilarity in request patterns or due to hardware failures and recovery. To quantify the effect of request desynchronization or file availability on the finish time for each (node, remote file) pair, we used the experimental setting presented in Sec. V-B. We grouped the nodes into three sets, based on their upload capacities, with each set containing the five fastest, slowest and median nodes respectively. We repeated the experiment and delayed the requests for remote files and the availability of local files by nodes that belonged to each one of the aforementioned sets. The delay we introduced was a fraction of the maximum finish time in the Max-Min topology.

In Fig. 15 we plot the normalized maximum finish time in Max-Min topology with respect to the maximum finish time in the same topology when file requests are synchronized and all the files are immediately available. We vary the lag and the set of nodes that are delayed. When the set of slow nodes is delayed, the maximum finish time is increased even for relatively small lag. Moreover, we observed that all the nodes finished at around the same time because the last pieces to be downloaded are those served by the slow nodes (despite the fact that local rarest first algorithm helps the dissemination of the these fresh pieces). When the set of median or fast nodes is delayed, the maximum finish time increases only for relatively high lag (a quarter of the maximum finish time) and only for the slowest peers.

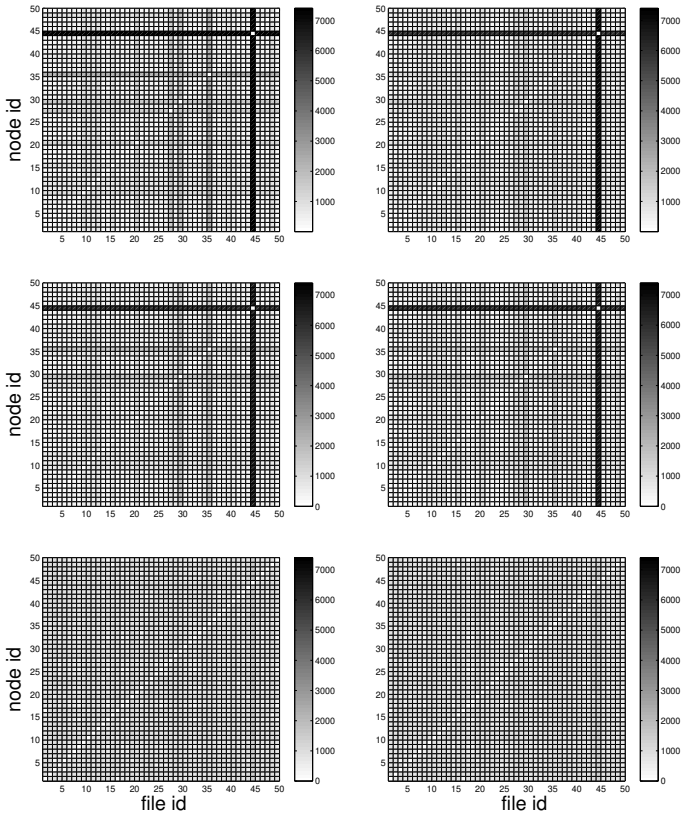

Fig. 14. Maximum finish time for all nodes and all files under different wirings (from left to right): MaxSum, Random, k-Widest, Rnd k-Widest, MaxMin strategies with cooperative nodes and Max-Min with upload-selfish slowest node, in the presence of a very slow node ($k/n = 0.04$).

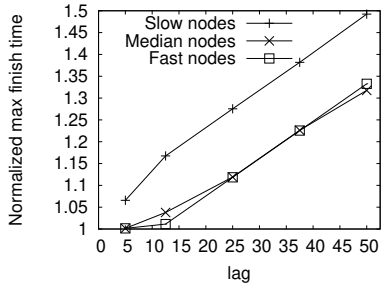

Fig. 15. Normalized maximum finish time when a set of nodes is delayed in Max-Min topology with respect to the maximum finish time when the all the file requests are synchronized and the file are immediately available. The lag is a percentage of the maximum finish time of the reference setting.

VIII. Conclusion

In this article we showed that swarming protocols for bulk data transfers perform much better when operating over optimized overlay topologies that take into consideration the end-to-end performance characteristics of the underlying network. Such topologies improve the aggregate transmission capacity of nodes, but where they make a huge difference compared to existing heuristic approaches, is on relieving bottleneck points. Random and myopic heuristics used in practice lack the required sophistication for overcoming such bottlenecks.

Our optimized topologies are formed in a distributed fashion and are oblivious to the details of the swarming protocol that runs on top. They leverage the available bandwidth of the underlying network and abstract the swarming protocol by viewing it as a series of max-flows. Thus they can benefit a variety of swarming protocols with different upload/download scheduling characteristics. Since our topologies are data-blind, it is the job of the swarming protocol to make the best use of the end-to-end bandwidth that they offer. To that end, we have shown that a commonly parameterized swarming protocol is far from being optimal.

References

[1] G. Smaragdakis, N. Laoutaris, P. Michiardi, A. Bestavros, J. W. Byers, and M. Roussopoulos, "Swarming on Optimized Graphs for n-way Broadcast," in *Proc. of IEEE INFOCOM '08*, Phoenix, AZ, 2008.

[2] B. Cohen, "Incentives Build Robustness in BitTorrent," in *Proc. of First Workshop on Economics of Peer-to-Peer Systems*, Berkeley, CA, 2003.

[3] X. Li, F. Bian, M. Crovella, C. Diot, R. Govindan, and G. Iannaccone, "Detection and Identification of Network Anomalies," in *Proc. of IMC '06*, Rio de Janeriro, Brazil, 2006.

[4] P. A. Bernstein and N. Goodman, "Concurrency Control in Distributed Database Systems," *ACM Computing Surveys*, vol. 13, no. 2, 1981.

[5] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. H. Katz, A. Konwinski, G. Lee, D. A. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "Above the Clouds: A Berkeley View of Cloud Computing," EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2009-28, Feb 2009.

[6] A. Legout, N. Liogkas, E. Kohler, and L. Zhang, "Clustering and Sharing Incentives in BitTorrent Systems," in *Proc. of ACM SIGMETRICS '07*, San Diego, CA, 2007.

[7] A. Fabrikant, A. Luthra, E. Maneva, C. H. Papadimitriou, and S. Shenker, "On a Network Creation Game," in *Proc. of ACM PODC '03*, Boston, MA, 2003.

[8] N. Laoutaris, G. Smaragdakis, A. Bestavros, and J. Byers, "Implications of Selfish Neighbor Selection in Overlay Networks," in *Proc. of IEEE INFOCOM '07*, Anchorage, AK, 2007.

[9] D. Qiu and R. Srikant, "Modeling and Performance Analysis of BitTorrent-like Peer-to-Peer Networks," in *Proc. of ACM SIGCOMM '04*, 2004.

[10] L. Massoulie and M. Vojnovic, "Coupon Replication Systems," in *Proc. of ACM SIGMETRICS '05*, Banff, Alberta, Canada, 2005.

[11] Y. Fernandess and D. Malkhi, "On Collaborative Content Distribution using Multi-message Gossip," *Journal of Parallel and Distributed Computing*, vol. 67, no. 12, pp. 1232–1239, 2007.

[12] R. Kumar and K. W. Ross, "Optimal Peer-Assisted File Distribution: Single and Multi-Class Problems," in *Proc. of IEEE Workshop on Hot Topics in Web Systems and Technologies (HOTWEB)*, Boston, MA, 2006.

[13] S. Sanghavi, B. Hajek, and L. Massoulie, "Gossiping With Multiple Messages," *IEEE Transactions on Information Theory*, vol. 53, no. 12, pp. 4640–4654, 2007.

[14] A. R. Bharambe, C. Herley, and V. N. Padmanabhan, "Analyzing and Improving a BitTorrent Networks Performance Mechanisms," in *Proc. of IEEE INFOCOM '06*, Barcelona, Spain, 2006.

[15] M. Izal, G. Urvoy-Keller, E. W. Biersack, P. A. Felber, A. A. Hamra, and L. Garces-Erice, "Dissecting BitTorrent: Five Months in a Torrent's Lifetime," in *Proc. of PAM '04*, Antibes Juan-les-Pins, France, 2004.

[16] A. Legout, G. Urvoy-Keller, and P. Michiardi, "Rarest first and choke algorithms are enough," in *Proc. of ACM IMC '06*, Rio de Janeiro, Brazil, 2006.

[17] L. Massoulie, A. Twigg, C. Gkantsidis, and P. Rodriguez, "Randomized Decentralized Broadcasting Algorithms," in *Proc. of IEEE INFOCOM '07*, Anchorage, AK, 2007.

[18] J. Edmonds, "Edge-disjoint branchings," in *Proc. of the 9th Courant Computer Science Symposium on Combinatorial Algorithms, Algorithmics Press*, 1972, pp. 91–96.

[19] C. Gkantsidis and P. Rodriguez, "Network Coding for Large Scale Content Distribution," in *Proc. of IEEE INFOCOM '05*, Miami, FL, 2005.

[20] B. Fan, J. C. S. Lui, and D.-M. Chiu, "The Design Trade-Offs of BitTorrent-Like File Sharing Protocols," *IEEE/ACM Transactions on Networking*, vol. 17, no. 2, pp. 365–376, 2009.

[21] L. Guo, S. Chen, Z. Xiao, E. Tan, X. Ding, and X. Zhang, "Measurements, Analysis, and Modeling of BitTorrent-like Systems," in *Proc. of ACM IMC '05*, Berkeley, CA, 2005.

[22] Y. Tian, D. Wu, and K.-W. Ng, "Analyzing Multiple File Downloading in BitTorrent," in *Proc. of ICPP '06*, Washington, DC, 2006.

[23] M. Castro, P. Druschel, A.-M. Kermarrec, A. Nandi, A. Rowstron, and A. Singh, "SplitStream: High-bandwidth Multicast in Cooperative Environments," in *Proc. of ACM SOSP '03*, Bolton Landing, NY, 2003.

[24] D. Kostic, A. Rodriguez, J. Albrecht, and A. Vahdat, "Bullet: High-bandwidth Data Dissemination Using an Overlay Mesh," in *Proc. of ACM SOSP '03*, Bolton Landing, NY, 2003.

[25] H. Zhang, G. Neglia, D. Towsley, and G. L. Presti, "On Unstructured File Sharing Networks," in *Proc. of IEEE INFOCOM '07*, Anchorage, AK, 2007.

[26] M. Jain and C. Dovrolis, "End-to-end Available Bandwidth: Measurement Methodology, Dynamics, and Relation with TCP throughput," *IEEE/ACM Transactions on Networking*, vol. 11, no. 4, pp. 537–549, 2003.

[27] S. M. Hedetniemi, S. T. Hedetniemi, and A. L. Liestman, "A survey of gossiping and broadcasting in communication networks," *Networks*, vol. 18, pp. 319–349, 1988.

[28] V. Ribeiro, R. Riedi, R. Baraniuk, J. Navratil, and L. Cottrell, "pathChirp: Efficient Available Bandwidth Estimation for Network Paths," in *Proc. of PAM '03*, La Jolla, CA, 2003.

[29] V. Arya, N. Garg, R. Khandekar, A. Meyerson, K. Munagala, and V. Pandit, "Local search heuristics for k-median and facility location problems," *SIAM Journal on Computing*, vol. 33, no. 3, pp. 544–562, 2004.

[30] R. H. Bisseling, *Parallel Scientific Computation: A Structured Approach using BSP and MPI*. Oxford University Press, 2004.

[31] N. Feamster, L. Gao, and J. Rexford, "How to Lease the Internet in Your Spare Time," *SIGCOMM Computer Communication Review*, vol. 37, no. 1, pp. 61–64, 2007.

[32] N. Spring, R. Mahajan, D. Wetherall, and T. Anderson, "Measuring ISP Topologies with Rocketfuel," *IEEE/ACM Transactions on Networking*, vol. 12, no. 1, pp. 2–16, 2004.

[33] R. Mahajan, N. Spring, D. Wetherall, and T. Anderson, "Inferring Link Weights using End-to-end Measurements," in *IMW '02: Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurment*, Marseille, France, 2002.

[34] A. Singh, M. Castro, P. Druschel, and A. Rowstron, "Defending Against Eclipse Attacks on Overlay Networks," in *Proc. of ACM SIGOPS European Workshop*, Leuven, Belgium, 2004.

APPENDIX A
NP-HARDNESS OF MAXIMIZING THE MINIMUM MAX-FLOW

Consider a node s that wants to select a set of neighbors σ from a network composed of m nodes $v_i \in V$, n nodes $u_j \in U$, and a single node t, so as to maximize its broadcast bandwidth defined to be its minimum max-flow to any destination, *i.e.*, $\Phi(s,\sigma) = \min_{x\in(V\cup U\cup\{t\})} MF(s,x,\sigma)$, where $MF(s,x,\sigma)$ denotes the max-flow from s to x under strategy σ. Node s can use $k < m$ links whose bandwidth is b_1 if the other end-point belongs to V, and $\epsilon \approx 0$ in any other case, implying that an optimal strategy σ for s must satisfy $\sigma \subset V, |\sigma| = k$. Each node v_i has directed links of bandwidth b_2 to a subset U_i of the nodes of U. Each node u_j has a link of bandwidth b_3 to t. Node t has links of bandwidth b_1 to all nodes of V and U (see Fig. 16 for an illustration). Link bandwidths obey:

$$b_1 \gg b_2 \gg b_3 \tag{1}$$

Let $\phi(s,X,\sigma) = \min_{x\in X} MF(s,x,\sigma)$ denote s's minimum max-flow to any node in the set X. Combining $k < m$ and (1), we get that under any σ, at least one node of V will get s's flow only indirectly through t, *i.e.*:

$$\phi(s,V,\sigma) = MF(s,t,\sigma) \tag{2}$$

The max-flow from s to u_j is equal to the max-flow from s to t, plus b_2 for each connected path $s \to v_i \to u_j$ under σ, minus the amount of flow that crosses the link from u_j to t in a max-flow from s to t under σ. Since this flow on the (u_j,t) link is either 0, or $b_3 < b_2$ when there's at least one connected path $s \to v_i \to u_j$ in σ, we get $MF(s,u_j,\sigma) \geq MF(s,t,\sigma)$, $\forall u_j \in U$, or equivalently:

$$\phi(s,U,\sigma) \geq MF(s,t,\sigma) \tag{3}$$

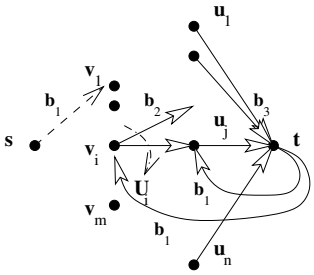

Fig. 16. Reduction from MAX-UNIQUES(k) to Max-Min.

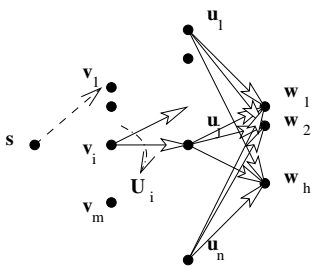

Fig. 17. Reduction from MAX-UNIQUES(k) to Max-Sum.

The max-flow to node t is:

$$MF(s,t) = b_3 \cdot paths(s,V,U) \tag{4}$$

where $paths(s,V,U)$ is the number of connected paths $s \to v_i \to u_j$ that do not share (v_i,u_j) edges, or equivalently the number of nodes u_j that carry a non-zero flow in a max-flow from s to t. Equations (2), (3) suggest that the maximization of the broadcast bandwidth calls for the maximization of $MF(s,t)$, which in view of (4), is achieved through the maximization of $paths(s,V,U)$. Maximizing $paths(s,V,U)$ requires choosing k subsets V_i so as to maximize the cardinality of their union. A straight-forward reduction from set-cover can be used to show that max $paths(s,V,U)$ is an NP-hard problem (see Appendix C). Therefore, maximizing the broadcast bandwidth is also NP-hard as it implies a solution to max $paths(s,V,U)$.

APPENDIX B
NP-HARDNESS OF MAXIMIZING THE SUM OF MAX-FLOWS

Consider a node s that wants to connect to a network composed of m nodes $v_i \in V$, n nodes $u_j \in U$, and h nodes $w_l \in W$, where h is a function of the out-degrees of the v_i's as will be explained shortly, so as to maximize the sum of its max-flows to all nodes in the union of V,U,W. Node s can use $k < m$ links whose bandwidth is 1 if the other end-point belongs to V, and $\epsilon \approx 0$ in any other case, implying than an optimal strategy σ for s must satisfy $\sigma \subset V, |\sigma| = k$. Each node v_i has directed links of unit bandwidth to a subset U_i of the nodes of U. Each node u_j has a link of unit bandwidth to each one of the nodes of W (see Fig. 17 for an illustration). The cardinality of W is equal to the highest out-degree of any node in V, *i.e.*, $h = \max_{1\leq i\leq m} |U_i|$.

Define $\psi(s,X,\sigma) = \sum_{x\in X} MF(s,x,\sigma)$ where $MF(s,x,\sigma)$ denotes the max-flow from s to x under strategy σ. Node s wants to select a strategy σ that maximizes $\Psi(s,\sigma) = \psi(s,V,\sigma) + \psi(s,U,\sigma) + \psi(s,W,\sigma)$ across all possible strategies. We will show that such an optimal strategy has to maximize the number of nodes in U to which there exists a connected path $s \to v_i \to u_j$.

Notice that $MF(s,v_i,\sigma) = 1$ iff $v_i \in \sigma$ and 0 otherwise, and thus $\psi(s,V,\sigma) = k$ independently of the particular strategy σ chosen. Therefore, we only need to care to maximize $\psi(s,U,\sigma)+\psi(s,W,\sigma)$. If s chooses to connect to v_i, meaning $v_i \in \sigma$, the contribution to $\psi(s,U,\sigma)$ will be $|U_i|$, because each outgoing link of v_i increases by 1 every max-flow from s to a node $u \in U_i$. The contribution to $\psi(s,W,\sigma)$ will be h for each node $u \in U$ that is reachable from s if

v_i is included in σ but becomes unreachable if it is taken out ("connecting" u increases all max-flows from s to nodes $w \in W$ by 1). Therefore if by switching $v_i \in \sigma$ with $v_{i'} \notin \sigma$ we get a strategy σ' which has a higher number of nodes of U reachable from s, then we should perform the switch because $\Psi(s, \sigma') > \Psi(s, \sigma)$. To see that, notice that the switch can hurt $\psi(s, U, \sigma)$ by at most $h - 1$, if v_i has the highest degree and $v_{i'}$ has degree 1 (it must have at least 1 to be increasing the number of unique u's reached), whereas it benefits $\psi(s, W, \sigma)$ by at least h as it increases the number of nodes of U reachable from s. The above argument implies that an optimal σ must maximize the number of unique nodes of U reachable from s. Therefore, an optimal solution to maximizing the sum of max-flows for s implies an optimal solution to the NP-hard problem MAX-UNIQUES(k) of Appendix C. Therefore, max sum max-flows is also an NP-hard problem.

APPENDIX C
NP-HARDNESS OF MAXIMIZING THE NUMBER OF UNIQUES

Let MAX-UNIQUES(k) be an optimization problem in which one has to select k subsets U_i, $1 \leq i \leq m$ of a set U with n elements so as to maximize the cardinality of the union $U(k) = \bigcup_{i \in choice} U_i$. Let UNIQUES$(k)$ be the corresponding decision problem in which one asks whether there is a *choice* leading to $|U(k)| = l$. UNIQUES(k) is clearly NP-complete because for $l = n$ a solution to UNIQUES(k) would imply a solution to SET-COVER. Therefore, MAX-UNIQUES(k) is NP-hard.

Georgios Smaragdakis received the Diploma in electronic and computer engineering from the Technical University of Crete, Greece, the Ph.D. degree in computer science from Boston University, MA, and he interned at Telefónica Research, Barcelona, Spain.

He is a Senior Research Scientist at Deutsche Telekom Laboratories and the Technical University of Berlin, Germany. His research interests include the design and analysis of computer networks and content distribution systems with main applications in overlay network creation and maintenance, service deployment, network storage, distributed caching, and network security.

Nikolaos Laoutaris is a researcher at the Internet research group of Telefonica Research in Barcelona. Prior to joining the Barcelona lab he was a postdoc fellow at Harvard University and a Marie Curie postdoc fellow at Boston University. He got his PhD in computer science from the University of Athens in 2004. His general research interests are on system, algorithmic, and performance evaluation aspects of computer networks and distributed systems with emphasis on content distribution, overlay networks, P2P, and multimedia communications.

Pietro Michiardi received his M.S. in Communication Systems in 1999 from EURECOM and his M.S. in Electrical Engineering from Politecnico di Torino in 2001. In December 2004 Pietro received his Ph.D. in Computer Science from Telecom ParisTech (former ENST, Paris). Since January 2005, Pietro is an Assistant Professor in the Networking & Security department at EURECOM. His research interests cover distributed algorithms and systems.

Azer Bestavros received the PhD degree in computer science from Harvard University in 1992. He is a professor in and former chairman of Computer Science at Boston University. His research interests are in networking and in real-time systems. His networking research contributions include his pioneering of the push content distribution model adopted years later by CDNs, his seminal work on traffic characterization and reference locality modeling, his work on various network transport, caching, and streaming media delivery protocols, his work on identifying and countering adversarial exploits of system dynamics, his work on game-theoretic approaches to overlay and P2P networking applications, his work on verification of network protocol compositions, and his work on virtualization services and programming environments for embedded sensor networks. His work has culminated so far in 13 PhD theses, more than 80 masters and undergraduate student projects, five US patents, two startup companies, and over 3,000 citations. His research has been funded by more than $15 million of government and industry grants. He has served as the general chair, program committee chair, officer, or PC member of most major conferences in networking and real-time systems. He has received distinguished service awards from both the ACM and the IEEE. He is the Chair of the IEEE Computer Society Technical Committee on the Internet and a distinguished speaker of the IEEE.

John W. Byers is an Associate Professor of Computer Science at Boston University. He is also Chief Scientist and a member of the Board of Directors at Adverplex, Inc., a quantitative search engine marketing firm based in Cambridge, MA. Dr. Byers joined B.U. in 1999 and has had an executive role at Adverplex since the company's founding in 2005. Prior to his appointment at B.U., he completed his Ph.D. in Computer Science at the University of California at Berkeley in 1997 and was a postdoctoral researcher at the International Computer Science Institute in Berkeley.

Professor Byers' research interests are broad and include algorithmic and economic aspects of networking, Internet content delivery, and pay-per-click advertising. Dr. Byers received the ACM SIGCOMM Test of Time Award in 2009 for his work on erasure-encoded content delivery, the IEEE ICDE Best Paper Award in 2004 for his work on sensor databases, and a National Science Foundation CAREER Award in 2001. He served terms on the editorial board of IEEE/ACM Transactions on Networking and on the executive committee of ACM SIGCOMM.

Mema Roussopoulos is an Assistant Professor of Computer Science at the University of Athens She served as an Assistant Professor of Computer Science at the University of Crete for two years and an Assistant Professor of Computer Science on the Gordon McKay Endowment at Harvard University for four years. Prior to Harvard, she was a Postdoctoral Fellow in the Computer Science Department at Stanford University. She received her PhD and Master's degrees in Computer Science from Stanford, and her Bachelor's degree in Computer Science from the University of Maryland at College Park. Her interests are in the areas of distributed systems, networking, mobile computing, and digital preservation. She is a recipient of the NSF CAREER Award and the Best Paper Award at ACM SOSP 2003.