

# Optimal Distributed P2P Streaming under Node Degree Bounds

Shaoquan Zhang, Ziyu Shao, Minghua Chen, and Libin Jiang

**Abstract**—We study the problem of maximizing the broadcast rate in peer-to-peer (P2P) systems under *node degree bounds*, i.e., the number of neighbors a node can simultaneously connect to is upper-bounded. The problem is critical for supporting high-quality video streaming in P2P systems, and is challenging due to its combinatorial nature. In this paper, we address this problem by providing the first distributed solution that achieves near-optimal broadcast rate under arbitrary node degree bounds, and over arbitrary overlay graph. It runs on individual nodes and utilizes only the measurement from their one-hop neighbors, making the solution easy to implement and adaptable to peer churn and network dynamics. Our solution consists of two distributed algorithms proposed in this paper that can be of independent interests: a network-coding based broadcasting algorithm that optimizes the broadcast rate given a topology, and a Markov-chain guided topology hopping algorithm that optimizes the topology. Our distributed broadcasting algorithm achieves the optimal broadcast rate over arbitrary P2P topology, while previously proposed distributed algorithms obtain optimality only for P2P complete graphs. We prove the optimality of our solution and its convergence to a neighborhood around the optimal equilibrium under noisy measurements or without time-scale separation assumptions. We demonstrate the effectiveness of our solution in simulations using uplink bandwidth statistics of Internet hosts.

**Index Terms**—P2P broadcasting, back pressure, network coding, Markov approximation, optimal and distributed algorithms

## I. INTRODUCTION

Peer-to-peer (P2P) systems have provided a scalable and cost effective way for streaming video in the past decade. Recent studies [11]–[14], however, indicate that the practical performance of P2P streaming systems can be far from their theoretical optimal.

There have been work studying the performance limit of P2P systems to understand and unleash their potential. One focus is on the *streaming capacity* problem [15] in P2P live streaming systems, i.e., maximizing the streaming rate subject to the peering and overlay topology constraints. The problem is critical for supporting high-quality video, which is determined by the streaming rate, in P2P live streaming systems. In this paper, we focus on the broadcast scenario where all peers in the system are receivers.

Shaoquan Zhang, Ziyu Shao and Minghua Chen are with the Department of Information Engineering, The Chinese University of Hong Kong, Shatin N.T., Hong Kong, email: zsq008@ie.cuhk.edu.hk; zyushao@ie.cuhk.edu.hk; minghua@ie.cuhk.edu.hk.

Libin Jiang was with California Institute of Technology, Pasadena, USA, email: lb.jiang@gmail.com.

This work was partially supported by the General Research Fund grants (Project No. 411008, 411209, 411010 and 411011) and an Area of Excellence Grant (Project No. AoE/E-02/08), all established under the University Grant Committee of the Hong Kong SAR, China. This work was also partially supported by two gift grants from Microsoft and Cisco.

The case of unconstrained peering on top of a complete graph is well studied, where the maximum broadcast rate is derived in several papers [1]–[3], [16], [17]. The case of unconstrained peering over general graph can also be addressed by using a centralized solution [5].

The streaming capacity problem becomes NP-Complete over general graph with *node degree bounds* [10]. Node degree is defined as the number of simultaneous active connections that a node maintains with its neighbors. Due to connection overhead costs, it is necessary to limit the number of simultaneous connections a peer can maintain. This naturally bounds the node degrees in P2P systems. For instance, in practical systems such as PPLive [18], the total number of neighbors of a node is usually bounded around 200, and the number of active neighbors of a node is usually bounded by 10-15 [15]. In such large P2P systems with hundreds of thousands of peers, the system topology is not a complete graph.

There has been work studying this challenging problem of maximizing streaming rate under node degree bounds and over general P2P graph. SplitStream/CoopNet [6], [7], ZIGZAG [8], PRIME [9] and most practical systems (such as PPLive [18] and UUSee [19]) bound node degree but do not provide rate optimality guarantee. Recently, the authors in [10] proposed a centralized Cluster-Tree algorithm that achieves near-optimal broadcast rate with high probability over complete graph, under the assumption that the node degree bound is at least logarithmic in the size of the network. A summary and comparison of previous work and this work are in Table I.

Despite of these exciting results, the following two important questions remain open:

- What is the maximum broadcast rate under arbitrary node degree bounds, and over general P2P overlay graph?
- How to achieve the maximum broadcast rate in a *distributed* manner?

Systems running distributed algorithms, compared with those running centralized algorithms, are more adaptable to peer churn and network dynamics.

In this paper, we answer the above two questions and make the following contributions:

- We provide the first distributed solution that achieves a broadcast rate arbitrarily close to the optimal under arbitrary node degree bounds, and over arbitrary overlay graph. Our solution runs on individual nodes and utilizes only the information from their one-hop neighbors.

Our solution consists of the following two algorithms that can be of independent interests.

- We propose a distributed broadcasting algorithm that achieves the optimal broadcast rate over arbitrary overlay

TABLE I  
SUMMARY AND COMPARISON OF PREVIOUS WORK AND THIS WORK FOR MAXIMIZING P2P BROADCAST RATE.

References	General Overlay Graph?	Arbitrary Node Degree Bound?	Exact or $1 - \epsilon$ Optimality?	Distributed Solution?
Mutualcast [1] and the algorithms in [2], [3]	×	×	✓	✓
Iterative in [4], [5]	✓	×	✓	×
CoopNet/SplitStream [6], [7]	×	✓	×	×
ZIGZAG [8], PRIME [9]	✓	✓	×	✓
Cluster-tree [10]	×	✓	conditionally optimal *	×
<b>This paper</b>	✓	✓	✓	✓

\* The Cluster-Tree algorithm is  $(1 - \epsilon)$ -optimal with high probability if the node degree bound is  $O(\log N)$ .

graph. Previous distributed P2P broadcasting algorithms are optimal only for complete overlay graph [1]–[3]. Our algorithm is based on network coding and utilizes back-pressure arguments.

- We also propose a distributed algorithm that optimizes the topology. In this algorithm, each node hops among their possible set of neighbors towards the best peering configuration. Our algorithm is inspired by a set of log-sum-exp approximation and Markov chain based arguments expounded in [20].
- We prove the optimality of the overall solution. We also prove its convergence to a neighborhood around the optimal equilibrium in the presence of noisy measurements or without time-scale separation assumptions. We demonstrate the effectiveness of our solution in simulations using uplink bandwidth statistics of Internet hosts.

## II. PROBLEM FORMULATION

### A. Settings and Notations

We model the P2P overlay network as a general directed graph  $G = (V, E)$ , where  $V$  denotes the set of nodes and  $E$  denotes the set of links. Each link in the graph corresponds to a TCP/UDP connection between two nodes. Let  $N_v$  denote the neighbor set of node  $v \in V$  in the graph. Each node  $v \in V$  is associated with an upload capacity  $C_v \geq 0$ . We assume there is no constraint on the downloading rate for each node  $v \in V$ . This assumption can be partly justified by the empirical observation that as residential broadband connections with asymmetric upload and download rates become increasingly dominant, bottlenecks typically are at the uplinks of the access networks rather than in the middle of the Internet.

As such, P2P networks have capacity limits on the nodes instead of links. This is different from traditional underlay networks where the capacity limits are on the links.

We focus on the single-source streaming scenario, i.e., a source  $s$  broadcasts a continuous stream of contents to the entire network. We denote the receiver set as  $R \triangleq V - \{s\}$ .

We consider the peering constraints that each node has a degree bound  $B_v$ , i.e., it can only exchange streaming content with up to a  $B_v$  number of neighbors *simultaneously* due to connection overhead cost. We allow different nodes to have different degree bounds. Fig. 1 shows four sample peering configurations of a 5-node network with node degree bound 3 for each node.

Let  $\mathcal{F}$  denote the set of all feasible peering configurations over graph  $G$  under node degree bounds. Given a configuration  $f \in \mathcal{F}$ , we obtain a connected sub-graph of  $G$  that satisfies

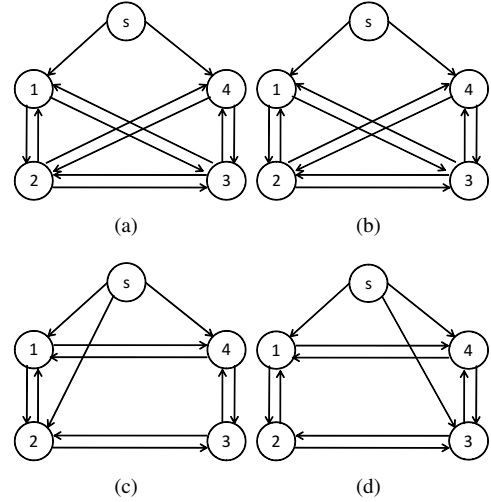


Fig. 1. Peering configuration examples for a 5-node network with node degree bound 3 for each node.

the node degree bound constraints. We denote this sub-graph as  $G_f = (V, E_f)$ , where  $E_f$  represents the set of links in this sub-graph. We denote  $N_{v,f}$  as the set of node  $v$ 's neighbors in this sub-graph. We have  $|N_{v,f}| \leq B_v$  where  $|\cdot|$  represents the size of a set.

### B. Problem Formulation and Our Approach

For a configuration  $f \in \mathcal{F}$ , let  $x_f$  be the maximum achievable broadcast rate under  $f$ , i.e., the highest rate at which every node in the system can receive the streaming content simultaneously. The problem of maximizing broadcast rate under node degree bounds can be formulated as follows:

$$\text{MRC} : \max_{f \in \mathcal{F}} x_f. \quad (1)$$

This problem is *combinatorial* in nature which is known to be NP-complete [10], and there is no efficient approximate solution to the problem even in a centralized manner.

In this paper, we address this problem by providing a distributed solution. In particular, we first develop a distributed broadcasting algorithm that can achieve  $x_f$  under arbitrary  $f \in \mathcal{F}$ . We then design a distributed algorithm that optimizes towards the best peering configurations. They operate in tandem to achieve a close-to-optimal broadcast rate under arbitrary node degree bounds, and over arbitrary overlay graph. We elaborate on these two algorithms in the following two sections.

## III. THE PROPOSED DISTRIBUTED BROADCASTING ALGORITHM

By exploiting network coding [21], we design a back-pressure based distributed broadcasting algorithm. Back-

pressure type algorithm is proposed initially in [22]. This type of algorithms select a subset of queues in the system with the maximum back-pressures and serve these queues subject to resource constraints, where back-pressure is defined as the difference between the queue at the local node and that of its downstream nodes. Back-pressure algorithm design has found applications in many network resource allocation domains [23], [24], [25]. In this paper, we apply this method for the first time to design distributed P2P broadcasting algorithm. Our algorithm can achieve the maximum broadcast rate over arbitrary P2P topology.

#### A. Routing vs. Network Coding

In P2P systems, there are two approaches for broadcasting contents: one is based on routing [26], in which nodes only store and forward packets; and the other is based on network coding [21], [26], in which a node is also allowed to mix information and output data as functions of the data it received. Some commercial P2P systems are built upon routing-based approach (e.g., PPLive [18]), and some are based on network coding (e.g., UUSee [19], [27])<sup>1</sup>. It is known that both routing and network coding approaches can achieve optimal broadcast rate over arbitrary P2P graph [2], [17]. Compared to routing-based approach, the network-coding based approach introduces additional packet header overhead for carrying coding coefficients (e.g., 3% extra overhead according to [29]) and computation complexity for encoding and decoding (e.g., [13], [27] discuss how to keep the complexity low). However, the network-coding based approach is robust to peer dynamics since there is no need for constructing and maintaining the spanning trees. In this section, we design a distributed broadcasting algorithm based on network coding that is robust to dynamics. In Section VII, we will discuss how the overall problem can be solved by using centralized solutions when only routing is allowed.

#### B. Network Coding Based Formulation

According to the Max-Flow-Min-Cut theorem, a data transmission of rate  $z$  between source  $s$  and a receiver  $d$  is *feasible* if and only if there exists a flow, denoted as  $f^d$ , satisfying the following flow conservation constraints:

$$\sum_{u \in \text{in}(v)} f_{uv}^d \leq \sum_{u \in \text{out}(v)} f_{vu}^d, \quad \forall v \in R - \{d\}, \quad (2)$$

$$z \leq \sum_{u \in \text{out}(s)} f_{su}^d, \quad (3)$$

$$0 \leq f^d, \quad (4)$$

where  $\text{in}(v) \triangleq \{u | (u, v) \in E_f\}$  is the set of nodes sending content to  $v$  under configuration  $f$ , and  $\text{out}(v) \triangleq \{u | (v, u) \in E_f, u \neq s\}$  is the set of nodes receiving content from  $v$ .

A powerful theorem established in [21] states that a multicast or broadcast rate  $z$  from  $s$  to a set of receivers is achievable if and only if  $z$  is feasible for  $s$  and any receiver  $d$ . This is a strong result as it says that if the network can support a

unicast rate of  $z$  between  $s$  and any receiver assuming other receivers' traffic is absent, then it can support a multicast rate of  $z$  to all the receivers simultaneously. Such rate  $z$  can be achieved by every node in the network performing network coding [21]. Further, authors in [29], [30] show that it is sufficient to perform random linear network coding.

In random linear network coding, by independently and randomly choosing a set of coding coefficients from a finite field, each node sends out the coded packet as a linear combination of the node's received packets. The combination information is specified by a *coefficient vector* in the packet header, which is updated by applying the same linear transformations as to the data. When one node receives a full set of linearly independent coded packets, it can decode and recover the original packets. In this paper, we focus on the distributed algorithm design. The discussions of decoding probability and implementation details can be found in [29], [30].

Under the setting of network coding, we can consider  $f^d$  as a "virtual" information flow between  $s$  and  $d$ . Multiple information flows "piggyback" together to transmit over the physical links. The actual physical rate over a physical link is only the maximum rate of individual information flows passing over it. Let  $g_{uv}$  be the physical flow rate over a link  $(u, v) \in E_f$ , then we have  $f_{uv}^d \leq g_{uv}$  for all  $d \in R$ .

With the above understanding, we formulate the problem of maximizing broadcast rate under configuration  $f$  as follows:

$$\mathbf{MP} : \max_{z, f, g \geq 0} U(z) \quad (5)$$

$$\text{s.t.} \quad \sum_{u \in \text{in}(v)} f_{uv}^d + z \mathbb{1}_{\{v=s\}} \leq \sum_{u \in \text{out}(v)} f_{vu}^d, \quad \forall v \in V - \{d\}, d \in R \quad (6)$$

$$f_{vu}^d \leq g_{vu}, \quad \forall v \in V, \forall u \in \text{out}(v), d \in R, \quad (7)$$

$$\sum_{u \in \text{out}(v)} g_{vu} \leq C_v, \quad \forall v \in V, \quad (8)$$

where  $U(z)$  is a twice-differentiable strictly concave utility function<sup>2</sup>,  $\mathbb{1}_{\{\cdot\}}$  denotes the indicator function. The constraints in (6) describe the flow conservation requirements. The constraints in (7) come from the piggybacking property of information flows. The node upload capacity constraints are in (8). The problem **MP** is a convex problem. All feasible broadcast rates must satisfy the constraints in (6)-(8) and are achievable by using random linear network coding.

#### C. Algorithm Design via Lagrange Decomposition

To proceed, we first relax the first set of constraints in (6) in problem **MP** to obtain a partial Lagrangian as follows:

$$\begin{aligned} L(z, f, g, \lambda) &= U(z) - \sum_{v \in V - \{d\}} \sum_{d \in R} \lambda_{v,d} \left( \sum_{u \in \text{in}(v)} f_{uv}^d + z \mathbb{1}_{\{v=s\}} - \sum_{u \in \text{out}(v)} f_{vu}^d \right) \\ &= U(z) - \sum_{v \in V} \sum_{d \in R} \lambda_{v,d} \left( \sum_{u \in \text{in}(v)} f_{uv}^d + z \mathbb{1}_{\{v=s\}} - \sum_{u \in \text{out}(v)} f_{vu}^d \right), \quad (9) \end{aligned}$$

where  $\lambda_{v,d}$ ,  $v \in V - \{d\}$ ,  $d \in R$  are Lagrange multipliers,  $\lambda_{d,d} = 0$ ,  $\forall d \in R$ , and  $\sum_{u \in \text{in}(s)} f_{us}^d = 0$ .

<sup>1</sup>We refer interested readers to [27], [28] for more details on performance of routing-based and network-coding-based practical P2P systems. We focus on optimal distributed P2P broadcasting algorithm design based on network coding in this paper.

<sup>2</sup>It might seem unnecessary to involve a strictly concave utility function in this formulation. The reason is that we later design a primal-dual algorithm to solve the problem, and using a strictly concave utility function can avoid its potential instability problem [17].

The strong duality holds for problem **MP** since the Slater conditions are satisfied [31]. Therefore, we can solve problem **MP** by finding the saddle points of  $L(z, \mathbf{f}, \mathbf{g}, \boldsymbol{\lambda})$ .

Noticing that

$$\sum_{v \in V} \sum_{d \in R} \lambda_{v,d} z \mathbb{1}_{\{v=s\}} = z \sum_{d \in R} \lambda_{s,d} \quad (10)$$

and

$$\sum_{v \in V} \sum_{d \in R} \lambda_{v,d} \left( \sum_{u \in \text{in}(v)} f_{uv}^d - \sum_{u \in \text{out}(v)} f_{vu}^d \right) = \sum_{d \in R} \sum_{v \in V} \sum_{u \in \text{out}(v)} f_{vu}^d (\lambda_{u,d} - \lambda_{v,d}), \quad (11)$$

we can find the saddle points of  $L(z, \mathbf{f}, \mathbf{g}, \boldsymbol{\lambda})$  by solving the following problem successively in  $z, \mathbf{f}, \mathbf{g}, \boldsymbol{\lambda}$ :

$$\min_{\lambda \geq 0} \left( \max_{z \geq 0} (U(z) - z \sum_{d \in R} \lambda_{s,d}) + \max_{\mathbf{f}, \mathbf{g} \geq 0} \sum_{d \in R} \sum_{v \in V} \sum_{u \in \text{out}(v)} f_{vu}^d (\lambda_{v,d} - \lambda_{u,d}) \right) \quad (12)$$

s.t. (7) – (8).

Given  $\boldsymbol{\lambda}$  and  $z$ , we consider the following scheduling subproblem on  $\mathbf{f}, \mathbf{g}$ :

$$\text{SSP} : \max_{\mathbf{f}, \mathbf{g} \geq 0} \sum_{d \in R} \sum_{v \in V} \sum_{u \in \text{out}(v)} f_{vu}^d (\lambda_{v,d} - \lambda_{u,d}) \quad (13)$$

s.t. (7) – (8).

The above linear programming problem has a structure that allows us to solve it distributedly. The first observation is that if an optimal  $\mathbf{g}^*$  is given, then an optimal  $\mathbf{f}^*$  can be obtained as follows:  $\forall u, v \in V, d \in R$ ,

$$(f_{vu}^d)^* = \begin{cases} 0, & \text{if } \lambda_{v,d} - \lambda_{u,d} \leq 0, \\ g_{vu}^*, & \text{otherwise.} \end{cases} \quad (14)$$

As such, it is sufficient to study the following problem in  $\mathbf{g}$ :

$$\max_{\mathbf{g} \geq 0} \sum_{v \in V} \sum_{u \in \text{out}(v)} g_{vu} w_{vu} \quad (15)$$

s.t.  $\sum_{u \in \text{out}(v)} g_{vu} \leq C_v, \forall v \in V$ ,

where

$$w_{vu} \triangleq \sum_{d \in R} [\lambda_{v,d} - \lambda_{u,d}]^+, \quad \forall (u, v) \in E_f. \quad (16)$$

denotes the aggregate back-pressure between two neighboring nodes  $u$  and  $v$ , and  $[\cdot]^+ \triangleq \max(\cdot, 0)$ .

For any  $v \in V$ , let

$$u^*(v) \triangleq \arg \max_{u \in \text{out}(v)} w_{vu} \quad (17)$$

be one of its neighbors with the maximum back-pressure (breaking ties arbitrarily). Then one optimal solution for problem **SSP** is as follows:

$$(g_{vu}^d)^* = \begin{cases} C_v, & \text{if } u = u^*(v), \\ 0, & \text{otherwise,} \end{cases} \quad (18)$$

and

$$(f_{vu}^d)^* = \begin{cases} 0, & \text{if } \lambda_{v,d} - \lambda_{u,d} \leq 0, \\ g_{vu}^*, & \text{otherwise.} \end{cases} \quad (19)$$

Given  $\mathbf{f}^*$  and  $\mathbf{g}^*$ , primal-dual algorithms can be designed to adapt  $z$  and  $\boldsymbol{\lambda}$  to pursue the desired optimal solution.

We summarize the above analysis into a distributed algorithm including the following components:

**Primal-dual Rate Control:** we pursue the saddle point in  $z$  and  $\boldsymbol{\lambda}$  simultaneously as follows:

$$\begin{cases} \dot{z} = \alpha [U'(z) - \sum_{d \in R} \lambda_{s,d}]_z^+, \\ \dot{\lambda}_{v,d} = k_{v,d} \left[ \sum_{u \in \text{in}(v)} (f_{uv}^d)^* + z \mathbb{1}_{v=s} - \sum_{u \in \text{out}(v)} (f_{vu}^d)^* \right]_{\lambda_{v,d}}^+, \\ \dot{\lambda}_{d,d} = \lambda_{d,d} = 0, \quad \forall d \in R, \end{cases} \quad (20)$$

where  $\alpha$  and  $k_{v,d}$  are positive step sizes, and the function

$$[b]_a^+ = \begin{cases} \max(0, b), & a \leq 0, \\ b, & a > 0. \end{cases}$$

**Neighbor Scheduling, Content Scheduling, and Network Coding:** Every node  $v \in V$  maintains a queue storing packets that are intended for  $d$ . Whenever a transmission opportunity arises, node  $v$  chooses one neighbor  $u^*(v)$  with the maximum back-pressure according to (17).

If  $w_{vu^*(v)} > 0$ , node  $v$  sends packets to  $u^*(v)$  at rate  $C_v$ . Every output packet is constructed as follows. Node  $v$  chooses one packet from the head of each queue of  $d$  if  $\lambda_{v,d} - \lambda_{u^*(v),d} > 0$ , and output one random linear combination of these heard-of-queue packets. If otherwise  $w_{vu^*(v)} \leq 0$  or there is no head-of-line packets to code, node  $v$  does nothing.

We have the following observations.

- The Lagrangian variable  $\lambda_{v,d}$  is proportional to the length of queue storing packets that are intended for receiver  $d$ . The back-pressure  $w_{vu}$  measures the aggregate difference in the queues of all  $d \in R$  between  $v$  and  $u$ . The larger the back-pressure is, the more desperate node  $u$  wants to receive data from  $v$ .
- Our algorithm can be implemented in a distributed manner. It only requires nodes to exchange information with its one-hop neighbors, and thus is robust to peer churn and system dynamics. When a new peer arrives, it connects to a set of neighbors, assigned by the streaming server or trackers. Then the peer starts exchanging streaming data with them following the strategy defined by our algorithm. When a peer leaves, its neighbors are informed and then close the connections. For the network coding operation, theoretically we need to adjust the size of field where the coding coefficients are chosen to make sure of the decoding probability when the number of nodes changes [32], [33]. While [29] and [13] show that in practice the finite field  $\mathbb{F}_{2^8}$  or  $\mathbb{F}_{2^{16}}$  is enough to have a sufficiently high decoding probability. Therefore, only local configuration changes corresponding to dynamics, which is easy to implement compared to centralized algorithms where typically global information is needed for whole configuration change (e.g., spanning trees reconstruction in spanning tree based solutions).
- Although our algorithm is designed for P2P broadcast scenarios, it also works for P2P multicast scenarios where helper nodes exist. The helper nodes simply also perform the operations described in (18)-(20). Our algorithm can be considered as the extension of the algorithm in [30], [34], [35] from link-capacity-limited underlay networks to node-capacity-limited overlay networks. The unique part of our algorithm is optimal neighbor scheduling which decides for each node how to allocate its node capacity

optimally among its neighbors for transmitting the coded traffic.

The following theorem characterizes the convergence of the proposed algorithm.

**Theorem 1:** The algorithm in (18)-(20) converges to the optimal solution of problem **MP** globally asymptotically in time.

The proof utilizes standard Lyapunov arguments and a Lyapunov function for primal-dual algorithm, similar to those used in [17], [36]. The proof is relegated to Appendix-A.

**Remark:** We derive our algorithm and prove its convergence based on a fluid model formulation. It is also possible to obtain a similar back-pressure based distributed algorithm with packet-level dynamics taken into account and prove its stability, following a set of Lyapunov drift arguments elaborated in [37].

#### IV. THE PROPOSED DISTRIBUTED TOPOLOGY HOPPING ALGORITHM

We recently proposed in [20] to use Markov chain as a principled approach in designing distributed algorithms for solving combinatorial network problems approximately. In particular, we show one can design distributed algorithms for a combinatorial network optimization problem in the following way. First, construct a special class of Markov chains with problem-specific steady-state distribution. Second, search for a Markov chain in this class that allows distributed implementation. If such Markov chain can be found, which is usually challenging and problem-specific, the distributed implementation directly yields a distributed algorithm for the problem.

In this paper, we follow the framework from [20] and design a distributed topology hopping algorithm for our problem (1). There are two steps in designing our algorithm under the Markov approximation framework [20]: log-sum-exp approximation and constructing problem-specific Markov chains that allows distributed implementation.

##### A. Log-Sum-Exp Approximation

First, the maximum broadcast rate can be approximated by a log-sum-exp function as follows:

$$\max_{f \in \mathcal{F}} x_f \approx \frac{1}{\beta} \log \left[ \sum_{f \in \mathcal{F}} \exp(\beta x_f) \right], \quad (21)$$

where  $\beta$  is a positive constant. Let  $|\mathcal{F}|$  denote the size of the set  $\mathcal{F}$ , then the approximation accuracy is known as follows [20]:

$$0 \leq \frac{1}{\beta} \log \left[ \sum_{f \in \mathcal{F}} \exp(\beta x_f) \right] - \max_{f \in \mathcal{F}} x_f \leq \frac{1}{\beta} \log |\mathcal{F}|. \quad (22)$$

As  $\beta$  approaches infinity, the approximation gap approaches zero. As discussed in [20], however, usually  $\beta$  should not take too large values as there are practical constraints or convergence rate concerns in the algorithm design afterwards.

To better understand the log-sum-exp approximation, we associate with each configuration  $f \in \mathcal{F}$  a probability  $p_f$ . Consider the following problem

$$\text{MRC} - \text{EQ} : \max_{p \geq 0} \sum_{f \in \mathcal{F}} p_f x_f \quad (23)$$

$$\text{s.t.} \quad \sum_{f \in \mathcal{F}} p_f = 1. \quad (24)$$

Its optimal value is  $\max_{f \in \mathcal{F}} x_f$  and is obtained by setting the probability corresponding to one of the best configurations to be one and the rest probabilities to be zero. Hence, problem **MRC** – **EQ** is equivalent to the original problem **MRC**.

On the other hand, according to [20] we have the following observations.

**Theorem 2 (cf. [20]):** The optimal value of the following optimization problem

$$\text{MRC} - \beta : \max_{p \geq 0} \sum_{f \in \mathcal{F}} p_f x_f - \frac{1}{\beta} \sum_{f \in \mathcal{F}} p_f \log p_f \quad (25)$$

$$\text{s.t.} \quad \sum_{f \in \mathcal{F}} p_f = 1 \quad (26)$$

is given by  $\frac{1}{\beta} \log \left[ \sum_{f \in \mathcal{F}} \exp(\beta x_f) \right]$ . The optimal solution of problem **MRC** –  $\beta$  is given by

$$p_f^*(\mathbf{x}) = \frac{\exp(\beta x_f)}{\sum_{f' \in \mathcal{F}} \exp(\beta x_{f'})}, \quad \forall f \in \mathcal{F}. \quad (27)$$

As such, by the log-sum-exp approximation in (21), we obtain an approximate version of the maximum broadcast rate problem **MRC**, off by an *entropy* term  $-\frac{1}{\beta} \sum_{f \in \mathcal{F}} p_f \log p_f$ . If we can time-share among different configurations according to the optimal solution  $p_f^*(\mathbf{x})$  in (27), then we can solve the problem **MRC** approximately and obtain a close-to-optimal broadcast rate.

##### B. Markov Chain Guided Algorithm Design

We design a Markov chain with state space being the set of all feasible peering configurations  $\mathcal{F}$  and stationary distribution as  $p_f^*(\mathbf{x})$  in (27). We implement the Markov chain to guide the system to optimize the configuration. As the system hops among configurations, the Markov chain converges and the configurations are time-shared according to the desired distribution  $p_f^*(\mathbf{x})$ .

The key lies in designing such Markov chain that allows distributed implementation. Since  $p_f^*(\mathbf{x})$  in (27) is product-form, it suffices to focus on designing time-reversible Markov chains [20].

Let  $f, f' \in \mathcal{F}$  be two states of Markov chain, and denote  $q_{f,f'}$  as the transition rate from state  $f$  to  $f'$ . We have two degrees of freedom in designing a time-reversible Markov chain:

- **The state space structure:** we can add or cut direct transitions between any two states, given that the state space remains connected and any two states are reachable from each other.
- **The transition rates:** we can explore various options in designing  $q_{f,f'}$ , given that the detailed balance equation is satisfied, i.e.,

$$p_f^*(\mathbf{x}) q_{f,f'} = p_{f'}^*(\mathbf{x}) q_{f',f}, \quad \forall f, f' \in \mathcal{F}. \quad (28)$$

Satisfying the above equations guarantees the designed Markov chain has the desired stationary distribution as in (27).

Recall that for a node  $v \in V$ , the set of its neighbors under configuration  $f$  is denoted by  $N_{v,f}$ . We call node in  $N_{v,f}$   $v$ 's in-use neighbor and node in  $N_v \setminus N_{v,f}$   $v$ 's not-in-use neighbor. For the ease of explanation, we further define  $\mathcal{N}_f$  as the set of

all the node-pairs under  $f$ , i.e.,  $\mathcal{N}_f = \{\{v, u\}, \forall v \in V, u \in N_{v,f}\}$ . Note we do not differentiate node pairs  $\{u, v\}$  and  $\{v, u\}$ . As an example, for the peering configuration  $f$  shown in Fig. 1(b),  $\mathcal{N}_f$  is given by  $\{\{s, 1\}, \{s, 2\}, \{s, 4\}, \{1, 2\}, \{1, 4\}, \{2, 3\}, \{3, 4\}\}$ .

In our Markov chain design, we first specify its state space structure as follows: we set the transition rate  $q_{f,f'}$  to be zero, unless  $f$  and  $f'$  satisfy that  $|\mathcal{N}_f \setminus \mathcal{N}_{f'}| = 1$  or  $|\mathcal{N}_{f'} \setminus \mathcal{N}_f| = 1$ . In other words, we only allow direct transitions between two configurations if such transitions correspond to a single node adding a new node in its in-use neighbor set or removing one in-use neighbor from its in-use neighbor set.

Second, given the state space structure of Markov chain, we design the transition rates to favor distributed implementation while satisfying the detailed balance equation in (28).

One possible option is to set  $q_{f,f'}$  to be  $\exp^{-1}(\beta x_f)$ . One way to implement this option is for every node to generate a timer according to its *measured* receiving rate and counts down accordingly. When the timer expires, the dedicated node performs the neighbor swapping and resets its timer. As simple as the implementation may sound, this option is expensive to implement. Once the peering configuration changes, the system needs to notify all the nodes to measure the new receiving rate and reset their timers accordingly. It is not clear how to implement such system-wide notification in a low-overhead manner.

In this paper, we design  $q_{f,f'}$  and  $q_{f',f}$  as follows:

$$q_{f,f'} = \frac{1}{\exp(\tau)} \frac{\exp(\beta x_{f'})}{\exp(\beta x_f) + \exp(\beta x_{f'})} \quad (29)$$

and

$$q_{f',f} = \frac{1}{\exp(\tau)} \frac{\exp(\beta x_f)}{\exp(\beta x_f) + \exp(\beta x_{f'})}, \quad (30)$$

where  $\tau$  is a constant. It is straightforward to verify that detailed balance equation is satisfied. As will be clear in the next subsection, our choices of transition rates do not require coordination or notification among peers in its implementation.

### C. Distributed Implementation

One distributed implementation of our designed Markov chain is briefly described as follows.

- **Initialization:** Each peer  $v \in V$  randomly selects neighbors from its neighbor list  $N_v$  under the node degree bound and builds connections with these selected neighbors.
- **Step 1:** Let  $f$  denote the current configuration. Each node  $v \in V$  generates an exponentially distributed random number independently with mean  $\frac{2\exp(\tau)}{|N_v|}$ , and counts down according to this number.
- **Step 2:** When the count-down expires, node  $v$  measures its current receiving rate as an estimate of the broadcast rate  $x_f$ . Then with probability  $\frac{|N_{v,f}|}{|N_v|}$  node  $v$  goes to the **Step 2a**; with probability  $\frac{|N_v| - |N_{v,f}|}{|N_v|}$ , node  $v$  goes to the **Step 2b**;
  - **Step 2a:** Node  $v$  randomly selects one in-use neighbor in  $N_{v,f}$  and removes it from  $N_{v,f}$ . Under the new peering configuration  $f'$ , node  $v$  measures its receiving rate as an estimate of  $x_{f'}$ . With the estimates of  $x_f$  and  $x_{f'}$ , peer  $v$  stays in the new configuration  $f'$

with probability  $\frac{\exp(\beta x_{f'})}{\exp(\beta x_f) + \exp(\beta x_{f'})}$ , and switches back to  $f$  with probability  $1 - \frac{\exp(\beta x_{f'})}{\exp(\beta x_f) + \exp(\beta x_{f'})}$ . Node  $v$  then repeats **Step 1**.

- **Step 2b:** Node  $v$  randomly selects one not-in-use neighbor in  $N_v \setminus N_{v,f}$ . If the node degree of the selected not-in-use node is equal to the bound or  $v$ 's node degree is equal to the bound, node  $v$  jumps back to **Step 1** immediately. Otherwise, node  $v$  adds this selected node into  $N_{v,f}$ . Under the new peering configuration  $f'$ , node  $v$  measures its receiving rate as an estimate of  $x_{f'}$ . With the estimates of  $x_f$  and  $x_{f'}$ , peer  $v$  stays in the new configuration  $f'$  with probability  $\frac{\exp(\beta x_{f'})}{\exp(\beta x_f) + \exp(\beta x_{f'})}$ , and switches back to  $f$  with probability  $1 - \frac{\exp(\beta x_{f'})}{\exp(\beta x_f) + \exp(\beta x_{f'})}$ . Node  $v$  then repeats **Step 1**.

It is straightforward to summarize the above implementation into a distributed algorithm that runs on individual nodes and utilizes only the measurement from their one-hop neighbors. The correctness of the implementation is shown as follows:

**Proposition 1:** The implementation in fact realizes a time-reversible Markov chain with stationary distribution in (27). The proof is relegated to Appendix-B.

**Remarks:** a) In **Step 1**, the generation of count-down timers does not depend on the receiving rate, thus the system does not need to notify the nodes about changes of peering configurations. b) With the above implementation, the system hops towards configurations with better broadcast rate probabilistically. For example, if  $x_{f'} > x_f$ , then the system will be more likely to stay in configuration  $f'$  than in  $f$ , and vice versa. c) With large values of  $\beta$ , the system hops towards better configurations more greedily. However, this may as well lead to the system getting trapped in locally optimal configurations. Hence there is a trade-off to consider when setting the value of  $\beta$ . Moreover, the value of  $\beta$  also affects the convergence rate of the time-reversible Markov chain to the desired stationary distribution. It is worth future investigation to further understand the impact of  $\beta$ . d) In the presence of peer dynamics, our algorithm incurs only simple actions based on local information. When a new peer arrives, a neighbor set and a neighbor list are assigned to it. The peer builds connections with the nodes in the neighbor set. Then the peer starts counting down as **Step 1** and follows the strategy of our algorithm. When a peer leaves, we just eliminate it from the neighbor list of its previous neighbors and end up connections.

## V. CONVERGENCE PROPERTIES OF OVERALL SOLUTION

We have designed the distributed broadcasting algorithm in Section III and the Markov chain guided topology hopping algorithm in Section IV. The pseudocodes of each algorithm are shown in Algorithm 1 and Algorithm 2 respectively. Both algorithms are simple to implement, run on each individual node, and only require nodes to exchange information with their neighbors.

If the broadcasting algorithm converges instantaneously, i.e., time-scale separation assumption holds, then we can obtain the accurate value of  $x_f$  for any configuration  $f \in \mathcal{F}$ . Transiting based on the accurate  $x_f$ , the designed Markov chain will

**Algorithm 1** Broadcasting Algorithm

---

```

1: The following procedure runs on each individual node
   independently.
2: For the source  $s$  and each time slot,
3:    $x \leftarrow \left[ x + \alpha(U'(x) - \sum_{d \in R} \lambda_{s,d}) \right]^+$ 
4: For each node  $v \in V$  and each time slot,
5:    $w^* \leftarrow 0$ 
6:   for  $u \in \text{out}(v)$  do
7:     for  $d \in R$  do
8:        $w_{vu} \leftarrow w_{vu} + \max(\lambda_{v,d} - \lambda_{u,d}, 0)$ 
9:     end for
10:    if  $w_{vu} > w^*$  then
11:       $w^* \leftarrow w_{vu}$ 
12:       $u^* \leftarrow u$ 
13:    end if
14:  end for
15:  if  $w_{vu^*} > 0$  then
16:    for  $d \in R$  do
17:      if  $\lambda_{v,d} - \lambda_{u^*,d} > 0$  then
18:         $f_{vu^*}^d \leftarrow C_v$ 
19:      end if
20:    end for
21:  end if
22:  for  $d \in R$  do
23:     $\lambda_{v,d} \leftarrow \left[ \lambda_{v,d} + k_{v,d}(\sum_{u \in \text{in}(v)} f_{uv}^d - \sum_{u \in \text{out}(v)} f_{vu}^d) \right]^+$ 
24:  end for

```

---

converge to the desired stationary distribution (27). Hence by operating these two algorithms in tandem, we obtain a close-to-optimal broadcast rate under arbitrary node degree bounds, and over arbitrary overlay graph. The optimality gap is characterized in (22).

In practice, however, it is possible to obtain only an inaccurate measurement or estimate of  $x_f$ . These inaccuracies root in two sources. One is the noisy measurements of the maximum broadcast rates given the configuration. The other is the fast state transition of Markov chain, i.e., the Markov chain transits before the underlying broadcasting algorithm converges and thus it transits based on inaccurate observations of the broadcast rates.

Consequently, the topology hopping Markov chain may *not* converge to the desired stationary distribution  $p_f^*(\mathbf{x})$ . This observation motivates our following study on the convergence of Markov chain in the presence of inaccurate transition rates.

For each configuration  $f \in \mathcal{F}$  with broadcast rate  $x_f$ , we assume that its corresponding inaccurate observed rate belongs to the bounded region  $[-\Delta_f, \Delta_f]$ .  $\Delta_f$  is the inaccuracy bound and can be different for different  $f$ .

For easy explanation of our approach, we further assume that the observed broadcast rate for configuration  $f$  only takes one of the following  $2n_f + 1$  discrete values:

$$\left[ x_f - \Delta_f, \dots, x_f - \frac{1}{n_f} \Delta_f, x_f, x_f + \frac{1}{n_f} \Delta_f, \dots, x_f + \Delta_f \right],$$

where  $n_f$  is a positive constant. Further, with probability  $\eta_{j,f}$ , the observed broadcast rate takes value  $x_f + \frac{j}{n_f} \Delta_f$ ,  $\forall j \in \{-n_f, \dots, n_f\}$  and  $\sum_{j=-n_f}^{n_f} \eta_{j,f} = 1$ .

With the inaccurate observed broadcast rates, the topology hopping behaves as follows. Suppose the current configuration

**Algorithm 2** Topology Hopping Algorithm

---

```

1: The following procedure runs on each individual node
   independently. We focus on a particular node  $v \in V$ .
2: procedure Initialization
   • Initialize  $N_v, B_v$ ; randomly connects to peers from  $N_v$ 
     under the degree bound.
   • Generate a timer that follows exponential distribution
     with mean equal to  $2 \exp(\tau) / (|N_v|)$  and begin counting
     down.
3: end procedure
4:
5: When the timer expires, invoke the procedure Transition.
6: procedure Transition
7:   With probability  $\frac{|N_{v,f}|}{|N_v|}$ ,
8:      $N_o \leftarrow N_{v,f}$ ;
9:     randomly remove one in-use neighbor from  $N_{v,f}$ ;
10:     $x_{f'} \leftarrow \sum_{u \in \text{in}(v)} f_{uv}^v$ ;
11:     $N_{v,f'} \leftarrow N_o$  with probability
        $1 - \exp(\beta x_{f'}) / (\exp(\beta x_f) + \exp(\beta x_{f'}))$ ;
12:    refresh the timer and begin counting down;
13:   With probability  $1 - \frac{|N_{v,f}|}{|N_v|}$ ,
14:      $N_o \leftarrow N_{v,f}$ ;
15:     randomly add one not-in-use neighbor  $v'$  in
        $N_v \setminus N_{v,f}$  to  $N_{v,f}$ ;
16:     if  $|N_{v,f}| = B_v$  or  $|N_{v',f}| = B_{v'}$ 
17:       refresh the timer and begin counting down;
18:     end if
19:      $x_{f'} \leftarrow \sum_{u \in \text{in}(v)} f_{uv}^v$ ;
20:      $N_{v,f'} \leftarrow N_o$  with probability
        $1 - \exp(\beta x_{f'}) / (\exp(\beta x_f) + \exp(\beta x_{f'}))$ ;
21:     refresh the timer and begin counting down;
22: end procedure

```

---

is  $f$  and the observed broadcast rate is  $x_f + \frac{j}{n_f} \Delta_f$ , where  $j \in \{-n_f, \dots, n_f\}$ . After some count-down process, the system hops to a new configuration  $f'$  and probes its broadcast rate. In configuration  $f'$ , the broadcast rate is observed as  $x_{f'} + \frac{j'}{n_{f'}} \Delta_{f'}$ ,  $j' \in \{-n_{f'}, \dots, n_{f'}\}$ . The system stays in the new configuration  $f'$  with probability

$$\frac{\exp(\beta(x_{f'} + \frac{j'}{n_{f'}} \Delta_{f'}))}{\exp(\beta(x_{f'} + \frac{j'}{n_{f'}} \Delta_{f'})) + \exp(\beta(x_f + \frac{j}{n_f} \Delta_f))},$$

and switches back to configuration  $f$  with probability

$$1 - \frac{\exp(\beta(x_{f'} + \frac{j'}{n_{f'}} \Delta_{f'}))}{\exp(\beta(x_{f'} + \frac{j'}{n_{f'}} \Delta_{f'})) + \exp(\beta(x_f + \frac{j}{n_f} \Delta_f))}.$$

By arguments similar to the proof of Proposition 1, the transition rate from configuration  $f$  with broadcast rate  $x_f + \frac{j}{n_f} \Delta_f$  to configuration  $f'$  with broadcast rate  $x_{f'} + \frac{j'}{n_{f'}} \Delta_{f'}$  is given by

$$\frac{\eta_{j',f'}}{\exp(\tau)} \cdot \frac{\exp(\beta(x_{f'} + \frac{j'}{n_{f'}} \Delta_{f'}))}{\exp(\beta(x_{f'} + \frac{j'}{n_{f'}} \Delta_{f'})) + \exp(\beta(x_f + \frac{j}{n_f} \Delta_f))}. \quad (31)$$

We construct a Markov chain to capture and study the above topology hopping behavior. In this Markov chain, a state is associated with a configuration and an observed broadcast rate. Given any configuration  $f \in \mathcal{F}$  and its corresponding  $x_f$ , there are  $2n_f + 1$  states in the extended Markov

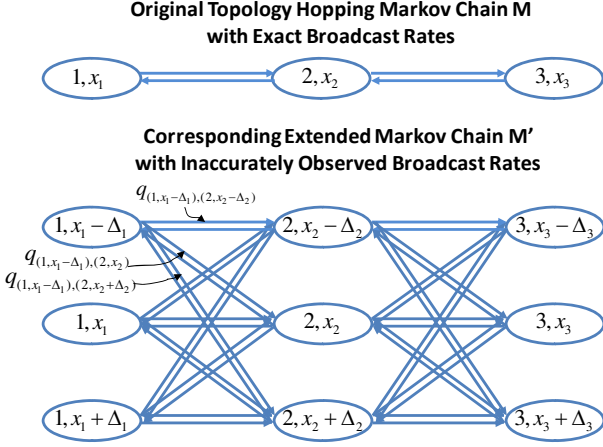


Fig. 2. An example of the original three-state topology hopping Markov chain and the extended Markov chain.  $M$  is the original topology hopping Markov chain with accurate broadcast rates.  $M'$  is the corresponding extended Markov chain with inaccurate broadcast rate observations. For each configuration  $f \in \{1, 2, 3\}$ , the observed broadcast rate takes values  $x_f - \Delta_f$ ,  $x_f$ ,  $x_f + \Delta_f$  with probability  $\eta_{-1,f}$ ,  $\eta_{0,f}$  and  $\eta_{1,f}$  respectively. The transition rates are assigned according to (32) and (33).

chain:  $(f, x_f + \frac{j}{n_f} \Delta_f)$ ,  $j \in \{-n_f, \dots, n_f\}$ . Further, Given direct transitions between configuration  $f$  and  $f'$  in the original topology hopping Markov chain, there are direct transitions between states  $(f, x_f + \frac{j}{n_f} \Delta_f)$  and  $(f', x_{f'} + \frac{j'}{n_{f'}} \Delta_{f'})$  ( $\forall j \in \{-n_f, \dots, n_f\}, j' \in \{-n_{f'}, \dots, n_{f'}\}$ ) in the corresponding new Markov chain. The corresponding transition rates are shown as follows:

$$q_{(f, x_f + \frac{j}{n_f} \Delta_f), (f', x_{f'} + \frac{j'}{n_{f'}} \Delta_{f'})} = \frac{\eta_{j', f'}}{\exp(\tau)} \cdot \frac{\exp(\beta(x_{f'} + \frac{j'}{n_{f'}} \Delta_{f'}))}{\exp(\beta(x_{f'} + \frac{j'}{n_{f'}} \Delta_{f'})) + \exp(\beta(x_f + \frac{j}{n_f} \Delta_f))} \quad (32)$$

and

$$q_{(f', x_{f'} + \frac{j'}{n_{f'}} \Delta_{f'}), (f, x_f + \frac{j}{n_f} \Delta_f)} = \frac{\eta_{j, f}}{\exp(\tau)} \cdot \frac{\exp(\beta(x_f + \frac{j}{n_f} \Delta_f))}{\exp(\beta(x_f + \frac{j}{n_f} \Delta_f)) + \exp(\beta(x_{f'} + \frac{j'}{n_{f'}} \Delta_{f'}))}, \quad (33)$$

where  $\sum_{j=-n_f}^{n_f} \eta_{j, f} = 1$  and  $\sum_{j'=-n_{f'}}^{n_{f'}} \eta_{j', f'} = 1$ . This new Markov chain can be thought as an extended version of the original topology hopping Markov chain. As an example, an extended Markov chain is shown and explained in Fig. 2.

The extended Markov chain has a unique stationary distribution since it is irreducible and only has a finite number of states. We can study the impact of inaccurate broadcast rates by comparing the stationary configuration distribution of the new Markov chain and that of the original topology hopping Markov chain.

We denote the stationary distribution of the *states* in the new Markov chain by

$$\tilde{\mathbf{p}} \triangleq [\tilde{p}_{f, x_f + \frac{j}{n_f} \Delta_f}, j \in \{-n_f, \dots, n_f\}, f \in \mathcal{F}]. \quad (34)$$

We also denote  $\bar{\mathbf{p}} : [\bar{p}_f(\mathbf{x}), f \in \mathcal{F}]$  as the stationary distribution of the *configurations* in the extended Markov chain. Given a configuration  $f \in \mathcal{F}$ , there are  $2n_f + 1$  states associated with  $f$  in the extended Markov chain. We have

$$\bar{p}_f(\mathbf{x}) = \sum_{j \in \{-n_f, \dots, n_f\}} \tilde{p}_{f, x_f + \frac{j}{n_f} \Delta_f}, \forall f \in \mathcal{F}. \quad (35)$$

Recall that the stationary distribution of the configurations for the original topology hopping Markov chain is  $\mathbf{p}^* : [p_f^*(\mathbf{x}), f \in \mathcal{F}]$ . We use the total variance distance [38] to quantify the difference between  $\mathbf{p}^*$  and  $\bar{\mathbf{p}}$ , as

$$d_{TV}(\mathbf{p}^*, \bar{\mathbf{p}}) \triangleq \frac{1}{2} \sum_{f \in \mathcal{F}} |p_f^* - \bar{p}_f|. \quad (36)$$

We have the following result:

**Theorem 3:** Let  $\Delta_{\max} = \max_{f \in \mathcal{F}} \Delta_f$ , and  $x_{\max} = \max_{f \in \mathcal{F}} x_f$ . The  $d_{TV}(\mathbf{p}^*, \bar{\mathbf{p}})$  are bounded as follows:

$$0 \leq d_{TV}(\mathbf{p}^*, \bar{\mathbf{p}}) \leq 1 - \exp(-2\beta\Delta_{\max}). \quad (37)$$

Further, the optimality gap in broadcast rates  $|\mathbf{p}^* \mathbf{x}^T - \bar{\mathbf{p}} \mathbf{x}^T|$  is bounded as below:

$$|\mathbf{p}^* \mathbf{x}^T - \bar{\mathbf{p}} \mathbf{x}^T| \leq 2x_{\max}(1 - \exp(-2\beta\Delta_{\max})). \quad (38)$$

The proof is relegated to Appendix-C.

**Remarks:** a) The upper bound on  $d_{TV}(\mathbf{p}^*, \bar{\mathbf{p}})$  shown in (37) is general, as it is independent of the number of configurations  $|\mathcal{F}|$ , the values of  $n_f$ , and the distributions of inaccurate observed rates  $\eta_{j, f}$  ( $-n_f \leq j \leq n_f, f \in \mathcal{F}$ ). b) The upper bound on  $d_{TV}(\mathbf{p}^*, \bar{\mathbf{p}})$  shown in (37) decreases exponentially with the worst inaccuracy bound  $\Delta_{\max}$  decreasing. c) It would be interesting to explore a tighter upper bound on  $d_{TV}(\mathbf{p}^*, \bar{\mathbf{p}})$  than the one in (37).

## VI. PERFORMANCE EVALUATION

We implement a packet-level simulator to our proposed solutions and use this simulator to evaluate the performance of our solutions.

### A. Settings

In our simulations, time is chopped into slots of equal length, and we adopt three different settings. In Setting I, we set the total number of nodes to be 100, and assign the node upload capacities randomly according to the distribution in Table II, which is obtained from the uplink bandwidth statistics of Internet hosts [39]. We set the source's upload capacity to be 768 kbps. With this upload capacity, the source is not the broadcast bottleneck [1], [3].

Setting II is the same as Setting I, except that we set the total number of nodes to be 10.

In Setting III, there are 4 different peering configurations as shown in Fig. 3. Every node has a unit capacity. Under configuration  $f_1$ ,  $f_2$  and  $f_3$  the maximum broadcast rate is 1, and under configuration  $f_4$  the maximum broadcast rate is 0.5.

When running our network coding based broadcasting algorithm, we set the updating step size of  $z$  and  $\lambda$  to be 0.1 and 0.00005 respectively. These parameters are empirically chosen to obtain smooth algorithm updating and small errors.

In our simulations, we assign node degree bounds in the following two ways. The first is to set identical bound on each node's node degree. The second is to set degree bound proportional to the node's upload capacity. This is based on the empirical observations that nodes with high upload capacities usually have more system resource (e.g., memory and CPU power) than nodes with low upload capacities. With more system resource, nodes can maintain more concurrent connections, thus have larger node degree bounds. In our second degree bounds assignment, nodes set their node degree bounds



TABLE II  
PEER UPLOAD CAPACITY DISTRIBUTION

Upload Capacity (kbps)	64	128	256	384	768
Fraction (%)	2.8	14.3	4.3	23.3	55.3

proportional to the ratio between their upload capacities and 64 kbps. In particular, nodes with 64 kbps have a degree bound of 2, and nodes with 128 kbps have a degree bound of 4, etc.

We carry out two sets of simulations. First, we evaluate the performance of our distributed broadcasting algorithm under Setting I and II. Second, we evaluate the overall performance when we combine the topology hopping algorithm and the broadcasting algorithm under Setting I and III. In these two sets of simulations, we also compare the performance under the two different degree bounds assignments.

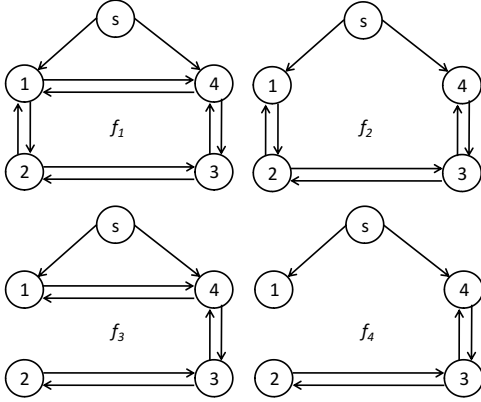


Fig. 3. Peering configurations under Setting III. For the ease of illustration, we only allow node 1 to add or remove neighbors between nodes 2 and 4. The rest nodes keep their neighbors fixed.

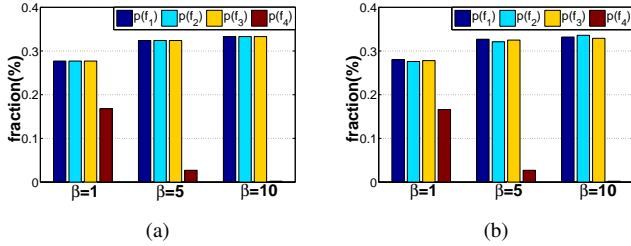


Fig. 5. a) Optimal configuration distribution for different values of  $\beta$  under Setting III; b) Configuration distribution obtained by our algorithm for different values of  $\beta$  under Setting III.

### B. Evaluation of the Proposed Broadcasting Algorithm

In this simulation, we evaluate our distributed broadcasting algorithm proposed in Section III. We randomly choose a sub-graph that satisfies the node degree bounds constraints, and run our algorithm over it. We evaluate three aspects of the proposed algorithm: 1) does it converge to the optimal broadcast rate as expected from the theoretical analysis? 2) How fast does it converge? 3) How would different values of degree bounds affect the maximum broadcast rate? The results are summarized in Fig. 4.

From Fig. 4(a) and Fig. 4(b), we see that our broadcasting algorithm converges. It converges faster in the small size network as shown in Fig. 4(a) than in the large size network

as shown in Fig. 4(b). From Fig. 4(d), we also see that the converged rate when the node degree bound is 10 is very close to a theoretical upper bound – the optimal broadcast rate without degree bounds computed according to [1], [17], [3]. This suggests that our algorithm converges to the optimal broadcast rate.

Under different degree bounds, the optimal broadcast rate varies. Fig. 4(d) shows that the optimal broadcast rate increases when we increase the node degree bounds. We plot the CDF of peer receiving rates (after the broadcasting algorithm converges) for the cases where degree bound is 4, 10, and proportional to the peer's upload capacity. It's seen that when the bound is 10, the obtained rate is close to the full-mesh rate, which suggests that we do not need a large degree bound to achieve close to the full-mesh rate.

### C. Evaluation of the Overall Solution

Our overall solution, which combines the Markov chain guided topology hopping algorithm and the back-pressure and network coding based broadcasting algorithm, achieves the near optimal broadcast rate under arbitrary node degree bound and over arbitrary overlay graph. To evaluate its performance, we generate a sub-graph randomly, run our algorithms on every node, and evaluate the achieved broadcast rate.

The topology hopping algorithm runs on top of the broadcasting algorithm. Under given topology, the broadcasting algorithm achieves the optimal broadcast rate. Nodes swap neighbors based on their observed receiving rate, thus changing the topology from time to time. In the simulation, we run the broadcasting algorithm long enough so that it converges before the topology transits according to the Markov chain. This way, the overall algorithm converges to the close-to-optimal broadcast rate.

In all simulations, we compare our overall algorithm with our back-pressure and network coding based broadcasting algorithm to illustrate the benefit of topology hopping, and with a simple heuristic algorithm introduced below to illustrate the benefit of our overall solution. Remind that no existing works solve the problem of streaming-rate maximization under general node degree bounds and over arbitrary topology we studied in this paper.

The simple heuristic algorithm we compare our overall algorithm against is also composed of two parts: routing-based broadcasting algorithm and random topology hopping algorithm. In routing-based broadcasting algorithm, each peer evenly allocates its upload capacity to its neighbors. Given the topology and capacity allocation, a centralized routing strategy (e.g. spanning trees based solution) is used to achieve the best broadcast rate the system can support. Similarly, the random topology hopping algorithm runs on the top of the broadcasting algorithm. Every peer maintains a timer. When the timer of one peer expires, the peer randomly drops one active neighbor which is exchanging data with it, and then selects one random candidate from its feasible neighbor list and starts to exchange data with it. By doing so, we actually allow nodes running the simple scheme to have a node degree beyond the bounds. This relaxation gives the simple scheme more degree of freedom to optimize its performance. Overall,

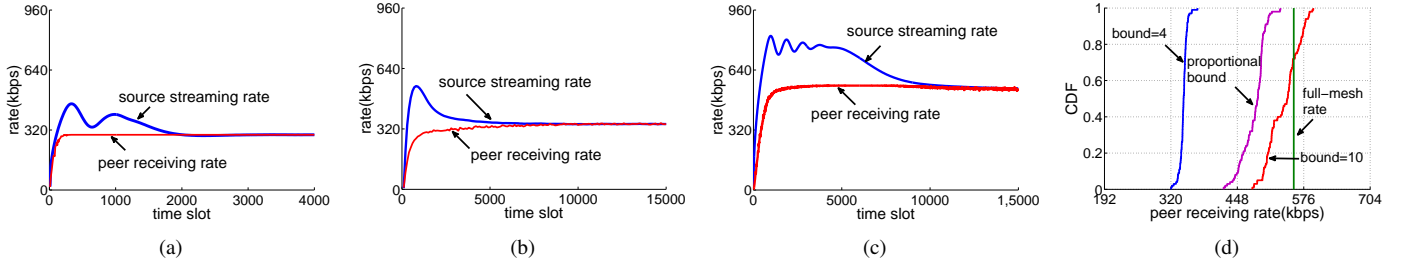


Fig. 4. Broadcasting algorithm evaluations. a) The source broadcast rate and average peer receiving rate under Setting II when degree bound is set to 3; b) The source broadcast rate and average peer receiving rate under Setting I when degree bound is set to 4; c) The source broadcast rate and average peer receiving rate under Setting I when degree bound is set to 10. d) This figure shows the impact of degree bound on the peer receiving rate under Setting I. The full-mesh rate is the maximum broadcast rate when the node degrees are unbounded [1].

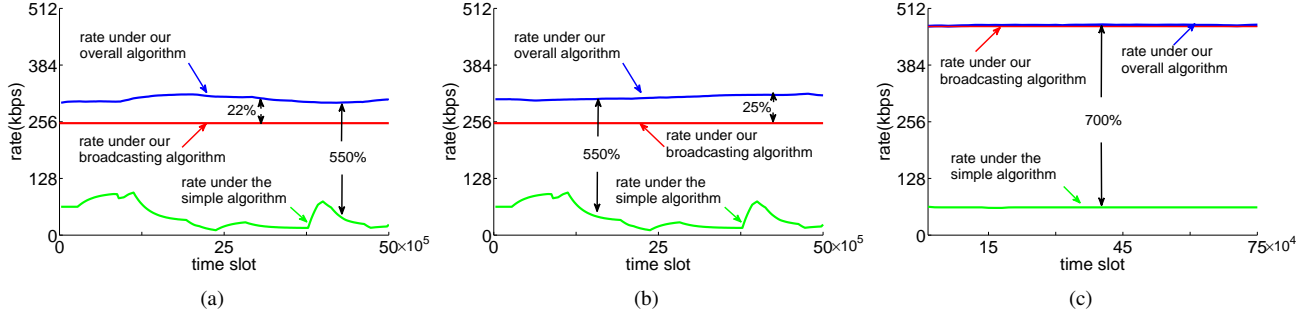


Fig. 6. Evaluation of our overall solution which combines the topology hopping algorithm and the broadcasting algorithm. a) The average peer receiving rate when the node degree bound is 3 and  $\beta$  is 20; b) The average peer receiving rate when the node degree bound is 3 and  $\beta$  is 50; c) The average peer receiving rate when peer degree bound is proportional to its upload capacity and  $\beta$  is 20. The percentage of average receiving rate improvement of our overall algorithm against our broadcasting algorithm and the simple heuristic algorithm are shown in these three figures. For example, in (a), 22% means that the average receiving rate of our overall algorithm is 1.22 times of that of our broadcasting algorithm, and 550% means that the average receiving rate of our overall algorithm is 6.5 times of that of the simple heuristic algorithm.

the topology changes randomly on the top under which peers use routing to exchange streaming data.

Our first observation is that our overall scheme converges to the solution that theory predicts. We carry out simulations under Setting III. Under this setting the optimal broadcast rate is 1. The optimal configuration solution to problem  $\text{MRC} - \beta$  is calculated and shown in Fig. 5(a) for different values of  $\beta$ . We run the overall scheme for this specific case and show the empirical configuration distribution in Fig. 5(b). Comparing the distributions in Fig. 5(a) and Fig. 5(b), we can see that the distribution obtained by our overall solution is very close to the optimal one. We also calculate the achieved broadcast rate under different values of  $\beta$ . For  $\beta = 1, 5$  and 10, the broadcast rate is 0.917, 0.987, and 0.998 respectively. We see that with large  $\beta$ , the achieved broadcast rate is close to the optimal value 1, as predicted by our analysis in Section IV.

Next, we evaluate our overall solution under Setting I. In Fig. 6(a) and Fig. 6(b), the broadcast rates obtained are 305 kbps and 312 kbps respectively. They are about 22% and 25% higher respectively than the broadcast rate 250 kbps achieved by running the broadcasting algorithm over a randomly chosen topology. This demonstrates the advantage of performing topology hopping to optimize the configuration, as compared to only randomly choosing topology.

By setting node degree bounds proportional to peers' upload capacity, nodes with higher upload capacity maintain more connections. From Fig. 6(c), we observe that this flexibility offers a broadcast rate of 475 kbps. Although the additional gain of topology hopping is small under the specific P2P

simulation settings (e.g., node uplink capacity distribution), we remark that our topology-hopping based algorithm is theoretically guaranteed to achieve close-to-optimal streaming rate under arbitrary node degree bounds and P2P settings, while the broadcasting algorithm with random topology selection has no performance guarantee. By Fig. 6(c), we want to show that there could be empirical benefit of allowing node degree bounds to be proportional to peers' upload capacities, and it might be an interesting future direction to fully investigate the potential of such option.

From Fig. 6(a), Fig. 6(b) and Fig. 6(c), we observe that the average receiving rate of our overall algorithm is about 5.5-7 times higher than that of the simple algorithm. And also we can see from Fig. 6(a) and Fig. 6(b), our algorithm can achieve smoother streaming rate than the simple algorithm because our algorithm optimizes the topology hopping and stays in the optimal topology while the simple algorithm hops among topologies randomly and arbitrarily.

## VII. DISCUSSIONS AND FUTURE WORK

In this paper, we propose a distributed solution to achieve a near-optimal broadcast rate under arbitrary node degree bounds, and over arbitrary overlay graph. Our solution is distributed and consists of two algorithms that can be of independent interests. The first is a distributed broadcasting algorithm that optimizes the broadcast rate given a P2P topology. It is derived from a network coding based problem formulation and utilizes back-pressure arguments. It can be

considered as the extension of the algorithm in [30] from link-capacity-limited underlay networks to node-capacity-limited overlay networks. The second algorithm is a Markov chain guided hopping algorithm that optimizes the topology, inspired by the Markov Approximation framework introduced in [20].

Assuming the underlying broadcasting algorithm converges instantaneously, the topology hopping algorithm converges to the optimal configuration distribution. When the broadcasting algorithm does not converge fast enough, the topology hopping Markov chain transits based on inaccurate observations of the maximum broadcast rates associated with the configurations. We show that the topology hopping algorithm still converges, but to a sub-optimal configuration distribution. We characterize an upper bound on the total variance distance between the optimal and sub-optimal configuration distributions, as well as an upper bound on the gap between the achieved and the optimal broadcast rates. We show that both bounds decrease exponentially as the bound on inaccuracy decreases.

Using uplink bandwidth statistics of Internet hosts, our simulations validate the effectiveness of the proposed solutions, and demonstrate the advantage of allowing node degree bounds to scale linearly with their upload capacities.

In the scenarios where network coding is not allowed, we can formulate the broadcasting problem in Subsection III-B as a linear program to construct a feasible node capacity allocation so that the sum of rate of all spanning trees is maximized [15], which is solvable by centralized LP algorithms. Then we can design the overall algorithm in the following way. The overall algorithm is composed of two separate algorithms: the spanning tree based broadcasting algorithm and the Markov chain guided hopping algorithm. The topology hopping algorithm is same as the one in Section IV which runs on the top of the broadcasting algorithm and guides the topology hopping. Compared to our distributed overall algorithm when network coding is applied, this algorithm is centralized making it unsuited for use in a dynamically changing systems.

Two interesting future directions are as follows. First, the convergence rate of our solution is determined by the mixing time of the topology-hopping Markov chain, which can be substantial for large P2P systems. It is thus of great interest to explore the design of topology-hopping Markov chains that mix fast and at the same time allows distributed implementation. Second, while our algorithms adapt well to peer dynamics, our theoretical analysis is for static scenarios. How to extend the analysis to dynamic scenarios such as those observed in practical P2P systems [40] is another interesting future direction.

## REFERENCES

- [1] J. Li, P. A. Chou, and C. Zhang, "Mutualcast: an efficient mechanism for content distribution in a p2p network," in *Proc. ACM SIGCOMM Asia Workshop*, 2005.
- [2] L. Massoulié, A. Twigg, G. Gkantsidis, and P. Rodriguez, "Randomized Decentralized Broadcasting Algorithms," in *Proc. IEEE INFOCOM*, 2007.
- [3] R. Kumar, Y. Liu, and K. Ross, "Stochastic fluid theory for p2p streaming systems," in *Proc. IEEE INFOCOM*, 2007.
- [4] Y. Cui, Y. Xue, and K. Nahrstedt, "Optimal resource allocation in overlay multicast," *IEEE Trans. Parallel and Distributed Systems*, 2006.
- [5] S. Sengupta, S. Liu, M. Chen, M. Chiang, J. Li, and P. A. Chou, "Peer-to-peer streaming capacity," *IEEE Trans. Information Theory*, 2011.
- [6] M. Castro, P. Druschel, A. Kermarrec, A. Nandi, A. Rowstron, and A. Singh, "SplitStream: high-bandwidth multicast in cooperative environments," in *Proc. ACM SOSP*, 2003.
- [7] V. Padmanabhan and K. Sripanidkulchai, "The case for cooperative networking," in *Proc. IPTPS*, 2002.
- [8] D. A. Tran, K. Hua, and T. Do, "ZIGZAG: An efficient peer-to-peer scheme for media streaming," in *Proc. IEEE INFOCOM*, 2003.
- [9] N. Magharei and R. Rejaie, "PRIME: Peer-to-peer receiver-driven mesh-based streaming," *IEEE/ACM Trans. Networking*, 2009.
- [10] S. Liu, M. Chen, S. Sengupta, M. Chiang, J. Li, and P. A. Chou, "Peer-to-peer streaming capacity under node degree bound," in *Proc. IEEE ICDCS*, 2010.
- [11] C. Feng, B. Li, and B. Li, "Understanding the performance gap between pull-based mesh streaming protocols and fundamental limits," in *Proc. IEEE INFOCOM*, 2009.
- [12] L. Abeni, C. Kiraly, and R. L. Cigno, "On the optimal scheduling of streaming applications in unstructured meshes," in *Proc. IFIP Networking*, 2009.
- [13] M. Wang and B. Li, "R2: Random Push with Random Network Coding in Live Peer-to-Peer Streaming," *IEEE Journal on Selected Areas in Communications*, 2007.
- [14] X. Zhang, J. Liu, B. Li, and T. Yum, "CoolStreaming/DONet: A data-driven overlay network for efficient live media streaming," in *Proc. IEEE INFOCOM*, 2005.
- [15] M. Chen, M. Chiang, P. A. Chou, J. Li, S. Liu, and S. Sengupta, "Peer-to-peer streaming capacity: Survey and recent results," in *Proc. Allerton Conference*, 2009.
- [16] D. M. Chiu, R. W. Yeung, J. Huang, and B. Fan, "Can network coding help in p2p networks?" in *Proc. IEEE NetCod*, 2006.
- [17] M. Chen, M. Ponc, S. Sengupta, J. Li, and P. Chou, "Utility maximization in peer-to-peer systems," in *Proc. ACM SIGMETRICS*, 2008.
- [18] PPLive. [Online]. Available: <http://www.pplive.com>
- [19] UUSEE. [Online]. Available: <http://www.uusee.com>
- [20] M. Chen, S. C. Liew, Z. Shao, and C. Kai, "Markov approximation for combinatorial network optimization," in *Proc. IEEE INFOCOM*, 2010.
- [21] R. Ahlswede, N. Cai, S.-Y. R. Li, and R. W. Yeung, "Network information flow," *IEEE Trans. Information Theory*, 2000.
- [22] L. Tassiulas and A. Ephremides, "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks," *IEEE Trans. Automatic Control*, 1992.
- [23] N. McKeown, V. Anantharam, and J. Walrand, "Achieving 100% throughput in an input-queued switch," in *Proc. IEEE INFOCOM*, 1996.
- [24] A. Eryilmaz, R. Srikant, and J. Perkins, "Stable scheduling policies for fading wireless channels," *IEEE/ACM Trans. on Networking*, 2005.
- [25] M. Neely, E. Modiano, and C. Rohrs, "Dynamic power allocation and routing for time-varying wireless networks," *IEEE Journal on Selected Areas in Communications*, 2005.
- [26] J. Liu, S. Rao, B. Li, and H. Zhang, "Opportunities and challenges of peer-to-peer internet video broadcast," in *Proc. IEEE*, 2008.
- [27] Z. Liu, C. Wu, B. Li, and S. Zhao, "Uusee: Large-scale operational on-demand streaming with random network coding," in *Proc. IEEE INFOCOM*, 2010.
- [28] X. Hei, C. Liang, J. Liang, Y. Liu, and K. Ross, "A Measurement Study of a Large-Scale P2P IPTV System," *IEEE Trans. Multimedia*, 2007.
- [29] P. Chou, Y. Wu, and K. Jain, "Practical network coding," in *Proc. Allerton Conference*, 2003.
- [30] T. Ho and H. Viswanathan, "Dynamic algorithms for multicast with intra-session network coding," *IEEE Trans. Information Theory*, 2005.
- [31] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.
- [32] T. Ho, M. Médard, R. Koetter, D. Karger, M. Effros, J. Shi, and B. Leong, "A random linear network coding approach to multicast," *IEEE Trans. Information Theory*, 2004.
- [33] P. Sanders, S. Egner, and L. Tolhuizen, "Polynomial time algorithms for network information flow," in *Proc. ACM Symposium on Parallel Algorithms and Architectures*, 2003.
- [34] T. Cui, L. Chen, and T. Ho, "On distributed scheduling in wireless networks exploiting broadcast and network coding," *IEEE Trans. Communications*, 2010.
- [35] L. Chen, T. Ho, M. Chiang, S. Low, and J. Doyle, "Congestion control for multicast flows with network coding," *IEEE Trans. Information Theory*, 2012.
- [36] R. Srikant, *The Mathematics of Internet Congestion Control*. Birkhäuser, 2004.
- [37] L. Georgiadis, M. Neely, and L. Tassiulas, *Resource allocation and cross layer control in wireless networks*. Now Pub, 2006.
- [38] P. Diaconis and D. Stroock, "Geometric bounds for eigenvalues of Markov chains," *The Annals of Applied Probability*, 1991.

- [39] C. Huang, J. Li, and K. W. Ross, "Can internet video-on-demand be profitable?" in *Proc. ACM SIGCOMM*, 2007.
- [40] F. Wang, J. Liu, and Y. Xiong, "Stable peers: Existence, importance, and application in peer-to-peer live video streaming," in *Proc. IEEE INFOCOM*, 2008.
- [41] S. Zhang, Z. Shao, and M. Chen, "Optimal distributed p2p streaming under node degree bounds," *Technical Report, Department of Information Engineering, The Chinese University of Hong Kong*, available at <http://arxiv.org/abs/1008.5248>.
- [42] F. Kelly, *Reversibility and stochastic networks*. Wiley, Chichester, 1979.

## APPENDIX

### A. Proof of Theorem 1

Due to space limit, the proof can be found in [41].

### B. Proof of Proposition 1

By two conditions for state space structure of Markov chain, we know that all configurations can reach each other within a finite number of transitions, thus the constructed Markov chain is irreducible. Further, it is a finite state ergodic Markov chain with a unique stationary distribution. We now show that the stationary distribution of the constructed Markov chain is indeed (27).

Now we verify that under the implementation, the state transition rate from  $f$  to  $f'$  satisfies (29).

In our Markov chain design, we only allow direct transitions between two configurations if such transitions correspond to a single node adding a new neighbor or removing a neighbor, i.e.,  $|\mathcal{N}_f \setminus \mathcal{N}_{f'}| = 1$  or  $|\mathcal{N}_{f'} \setminus \mathcal{N}_f| = 1$ . We consider these two scenarios separately in the following.

Let  $f \rightarrow f'$  denote the event that when the timer expires the process will enter state  $f'$  after leaving the current state  $f$ . The probability of this event is denoted by  $\Pr(f \rightarrow f')$ .

When  $|\mathcal{N}_f \setminus \mathcal{N}_{f'}| = 1$ , assuming  $\mathcal{N}_f \setminus \mathcal{N}_{f'} = (v, u)$ , the event  $f \rightarrow f'$  can be divided into two disjoint events: the event that node  $v$ 's timer expires, then node  $v$  selects node  $u$  to remove and removes it from its in-use neighbor set and the event that node  $u$ 's timer expires, then node  $u$  selects node  $v$  to remove and removes it from its in-use neighbor set. Denote these two events by  $f \xrightarrow{v-u} f'$  and  $f \xrightarrow{u-v} f'$ . Let  $v-u$  be the event that node  $v$  selects node  $u$  and removes it from its in-use neighbor set and  $u-v$  be the event that node  $u$  selects node  $v$  and removes it from its in-use neighbor set. Now we calculate the probability of  $f \xrightarrow{v-u} f'$  and  $f \xrightarrow{u-v} f'$  respectively.

$$\begin{aligned}
 & \Pr(f \xrightarrow{v-u} f') \\
 &= \Pr(v-u | v\text{'s timer expires}) \Pr(v\text{'s timer expires}) \\
 &= \frac{|\mathcal{N}_{v,f}|}{|\mathcal{N}_v|} \cdot \frac{1}{|\mathcal{N}_{v,f}|} \cdot \frac{\exp(\beta x_{f'})}{\exp(\beta x_f) + \exp(\beta x_{f'})} \cdot \frac{\frac{|\mathcal{N}_v|}{2 \exp(\tau)}}{\sum_{w \in V} \frac{|\mathcal{N}_w|}{2 \exp(\tau)}} \\
 &= \frac{1}{\sum_{v \in V} |\mathcal{N}_v|} \cdot \frac{\exp(\beta x_{f'})}{\exp(\beta x_f) + \exp(\beta x_{f'})} \quad (39)
 \end{aligned}$$

and

$$\begin{aligned}
 & \Pr(f \xrightarrow{u-v} f') \\
 &= \Pr(u-v | u\text{'s timer expires}) \Pr(u\text{'s timer expires}) \\
 &= \frac{|\mathcal{N}_{u,f}|}{|\mathcal{N}_u|} \cdot \frac{1}{|\mathcal{N}_{u,f}|} \cdot \frac{\exp(\beta x_{f'})}{\exp(\beta x_f) + \exp(\beta x_{f'})} \cdot \frac{\frac{|\mathcal{N}_u|}{2 \exp(\tau)}}{\sum_{w \in V} \frac{|\mathcal{N}_w|}{2 \exp(\tau)}} \\
 &= \frac{1}{\sum_{v \in V} |\mathcal{N}_v|} \cdot \frac{\exp(\beta x_{f'})}{\exp(\beta x_f) + \exp(\beta x_{f'})}. \quad (40)
 \end{aligned}$$

Therefore, we have

$$\begin{aligned}
 & \Pr(f \rightarrow f') \\
 &= \Pr(f \xrightarrow{v-u} f') + \Pr(f \xrightarrow{u-v} f') \\
 &= \frac{2}{\sum_{v \in V} |\mathcal{N}_v|} \cdot \frac{\exp(\beta x_{f'})}{\exp(\beta x_f) + \exp(\beta x_{f'})}. \quad (41)
 \end{aligned}$$

When  $|\mathcal{N}_{f'} \setminus \mathcal{N}_f| = 1$ , assuming  $\mathcal{N}_{f'} \setminus \mathcal{N}_f = (v, u)$ , similarly we divide  $f \rightarrow f'$  into two disjoint events  $f \xrightarrow{v+u} f'$  and  $f \xrightarrow{u+v} f'$ .  $f \xrightarrow{v+u} f'$  denotes the event that node  $v$ 's timer expires, then node  $v$  selects node  $u$  to add and adds it in its in-use neighbor set.  $f \xrightarrow{u+v} f'$  denotes the event that node  $u$ 's timer expires, then node  $u$  selects node  $v$  to add and adds it in its in-use neighbor set. Let  $v+u$  be the event that node  $v$  selects node  $u$  and adds it as one in-use neighbor and  $u+v$  be the event that node  $u$  selects node  $v$  and adds it as one in-use neighbor. Then we have

$$\begin{aligned}
 & \Pr(f \xrightarrow{v+u} f') \\
 &= \Pr(v+u | v\text{'s timer expires}) \Pr(v\text{'s timer expires}) \\
 &= \frac{|\mathcal{N}_v| - |\mathcal{N}_{v,f}|}{|\mathcal{N}_v|} \cdot \frac{1}{|\mathcal{N}_v| - |\mathcal{N}_{v,f}|} \cdot \frac{\exp(\beta x_{f'})}{\exp(\beta x_f) + \exp(\beta x_{f'})} \cdot \frac{\frac{|\mathcal{N}_v|}{2 \exp(\tau)}}{\sum_{w \in V} \frac{|\mathcal{N}_w|}{2 \exp(\tau)}} \\
 &= \frac{1}{\sum_{v \in V} |\mathcal{N}_v|} \cdot \frac{\exp(\beta x_{f'})}{\exp(\beta x_f) + \exp(\beta x_{f'})} \quad (42)
 \end{aligned}$$

and

$$\begin{aligned}
 & \Pr(f \xrightarrow{u+v} f') \\
 &= \Pr(u+v | u\text{'s timer expires}) \Pr(u\text{'s timer expires}) \\
 &= \frac{|\mathcal{N}_u| - |\mathcal{N}_{u,f}|}{|\mathcal{N}_u|} \cdot \frac{1}{|\mathcal{N}_u| - |\mathcal{N}_{u,f}|} \cdot \frac{\exp(\beta x_{f'})}{\exp(\beta x_f) + \exp(\beta x_{f'})} \cdot \frac{\frac{|\mathcal{N}_u|}{2 \exp(\tau)}}{\sum_{w \in V} \frac{|\mathcal{N}_w|}{2 \exp(\tau)}} \\
 &= \frac{1}{\sum_{v \in V} |\mathcal{N}_v|} \cdot \frac{\exp(\beta x_{f'})}{\exp(\beta x_f) + \exp(\beta x_{f'})}. \quad (43)
 \end{aligned}$$

Therefore, we have

$$\begin{aligned}
 & \Pr(f \rightarrow f') \\
 &= \Pr(f \xrightarrow{v+u} f') + \Pr(f \xrightarrow{u+v} f') \\
 &= \frac{2}{\sum_{v \in V} |\mathcal{N}_v|} \cdot \frac{\exp(\beta x_{f'})}{\exp(\beta x_f) + \exp(\beta x_{f'})}. \quad (44)
 \end{aligned}$$

In our implementation, under configuration  $f$ , peer  $v$  counts down with rate  $\frac{|\mathcal{N}_v|}{2} \exp^{-1}(\tau)$ . Therefore, the rate of leaving the state  $f$  is  $\sum_{v \in V} \frac{|\mathcal{N}_v|}{2} \exp^{-1}(\tau)$ . With the probability  $\Pr(f \rightarrow f')$ , the process jumps to state  $f'$  when leaving state  $f$ . So, the

transition rate from state  $f$  to  $f'$  is

$$\begin{aligned} q_{f,f'} &= \frac{2}{\sum_{v \in V} |N_v|} \cdot \frac{\exp(\beta \cdot x_{f'})}{\exp(\beta \cdot x_f) + \exp(\beta \cdot x_{f'})} \\ &\times \sum_{v \in V} \frac{|N_v|}{2} \exp^{-1}(\tau) \\ &= \exp^{-1}(\tau) \frac{\exp(\beta \cdot x_{f'})}{\exp(\beta \cdot x_f) + \exp(\beta \cdot x_{f'})}. \end{aligned} \quad (45)$$

With (27), we see that  $p_{f'}^*(\mathbf{x}) \cdot q_{f,f'} = p_f^*(\mathbf{x}) \cdot q_{f',f}$ ,  $\forall f, f' \in \mathcal{F}$ , i.e., the detailed balance equations hold. Thus the constructed Markov chain is time-reversible and its stationary distribution is indeed (27) according to Theorem 1.3 and Theorem 1.14 in [42].

### C. Proof of Theorem 3

We denote  $M$  as the original topology hopping Markov chain with exact broadcast rates, and  $M'$  as the corresponding extended Markov chain with inaccurately observed broadcast rates. For the convenience of expression, for all  $f \in \mathcal{F}$ ,  $j \in \{-n_f, \dots, n_f\}$ , we use  $f_j$  to represent the state  $(f, x_f + \frac{j}{n_f} \Delta_f)$  in the extended Markov chain  $M'$ , and  $\eta_{f_j}$  to represent distribution of inaccurate observed rates  $\eta_{f,j}$ .

Therefore, given direct transitions between configuration  $f$  and  $f'$  in the original topology hopping Markov chain  $M$ , there are direct transitions between states  $f_j$  and  $f'_k$  ( $\forall j \in \{-n_f, \dots, n_f\}, k \in \{-n_{f'}, \dots, n_{f'}\}$ ) in the extended Markov chain  $M'$ . Following (32) and (33), we have the corresponding transition rates

$$q_{f_j, f'_k} = \frac{\eta_{f'_k}}{\exp(\tau)} \cdot \frac{\exp(\beta(x_{f'} + \frac{k}{n_{f'}} \Delta_{f'}))}{\exp(\beta(x_{f'} + \frac{k}{n_{f'}} \Delta_{f'})) + \exp(\beta(x_f + \frac{j}{n_f} \Delta_f))} \quad (46)$$

and

$$q_{f'_k, f_j} = \frac{\eta_{f_j}}{\exp(\tau)} \cdot \frac{\exp(\beta(x_f + \frac{j}{n_f} \Delta_f))}{\exp(\beta(x_f + \frac{j}{n_f} \Delta_f)) + \exp(\beta(x_{f'} + \frac{k}{n_{f'}} \Delta_{f'}))}, \quad (47)$$

where  $\sum_{j=-n_f}^{n_f} \eta_{f_j} = 1$  and  $\sum_{k=-n_{f'}}^{n_{f'}} \eta_{f'_k} = 1$ .

Now we compute the stationary distribution of states for the extended Markov chain  $M'$ . By detailed balance equation, we have

$$p_{f_j} q_{f_j, f'_k} = p_{f'_k} q_{f'_k, f_j}, \quad \forall j \in \{-n_f, \dots, n_f\}, k \in \{-n_{f'}, \dots, n_{f'}\}. \quad (48)$$

Then we have

$$p_{f_j} \cdot \frac{1}{\eta_{f_j} \cdot \exp(\beta(x_f + \frac{j}{n_f} \Delta_f))} = p_{f'_k} \cdot \frac{1}{\eta_{f'_k} \cdot \exp(\beta(x_{f'} + \frac{k}{n_{f'}} \Delta_{f'}))}, \quad (49)$$

$$\forall j \in \{-n_f, \dots, n_f\}, k \in \{-n_{f'}, \dots, n_{f'}\}.$$

Therefore,

$$\frac{p_{f_0}}{\eta_{f_0} \cdot \exp(\beta x_f)} = \frac{p_{f'_0}}{\eta_{f'_0} \cdot \exp(\beta x_{f'})} \quad (50)$$

and

$$\frac{p_{f'_k}}{p_{f'_0}} = \frac{\eta_{f'_k}}{\eta_{f'_0}} \cdot \exp(\beta \frac{k}{n_{f'}} \Delta_{f'}), \quad \forall k \in \{-n_{f'}, \dots, n_{f'}\}. \quad (51)$$

Consider an arbitrary state  $\hat{f}_0$  in the extended Markov chain  $M'$ , where  $\hat{f} \in \mathcal{F}$  and  $\hat{f} \neq f, f'$ . Since state space of  $M'$  is

connected, we can always find a path to connect  $\hat{f}_0$  and  $f_0$  through a series of adjacent states  $\tilde{f}(1)_0, \dots, \tilde{f}(L)_0$ , and  $f_0 = \tilde{f}(1)_0, \tilde{f}(L)_0 = \hat{f}_0$ . Therefore,

$$\frac{p_{\hat{f}_0}}{p_{f_0}} = \prod_{l=1}^{L-1} \frac{p_{\tilde{f}(l+1)_0}}{p_{\tilde{f}(l)_0}} \quad (52)$$

and by (50) we have

$$\frac{p_{\tilde{f}(l+1)_0}}{\eta_{\tilde{f}(l+1)_0} \cdot \exp(\beta x_{\tilde{f}(l+1)})} = \frac{p_{\tilde{f}(l)_0}}{\eta_{\tilde{f}(l)_0} \cdot \exp(\beta x_{\tilde{f}(l)})}. \quad (53)$$

Then

$$\frac{p_{\hat{f}_0}}{\eta_{\hat{f}_0} \cdot \exp(\beta x_{\hat{f}})} = \frac{p_{f_0}}{\eta_{f_0} \cdot \exp(\beta x_f)}. \quad (54)$$

By (51) and (54), we know that  $\forall f \in \mathcal{F}$ ,

$$\frac{p_{f_0}}{\eta_{f_0} \cdot \exp(\beta x_f)} \text{ is a constant} \quad (55)$$

and

$$\frac{p_{f_j}}{p_{f_0}} = \frac{\eta_{f_j}}{\eta_{f_0}} \cdot \exp(\beta \frac{j}{n_f} \Delta_f), \quad \forall j \in \{-n_f, \dots, n_f\}. \quad (56)$$

On the other hand, we have

$$\sum_{f \in \mathcal{F}} \sum_{j=-n_f}^{n_f} p_{f_j} = 1. \quad (57)$$

By (55), (56) and (57), we obtain the stationary distribution of states for the extended Markov chain  $M'$  as follows:

$$\begin{aligned} \forall f \in \mathcal{F}, j \in \{-n_f, \dots, n_f\}, \\ \tilde{p}_{f_j} = \frac{\eta_{f_j} \cdot \exp(\beta(x_f + \frac{j}{n_f} \Delta_f))}{\sum_{f' \in \mathcal{F}} \sum_{k=-n_{f'}}^{n_{f'}} \eta_{f'_k} \cdot \exp(\beta(x_{f'} + \frac{k}{n_{f'}} \Delta_{f'}))}. \end{aligned} \quad (58)$$

The stationary distribution of peer configurations in the extended Markov chain  $M'$  is the probability distribution of aggregate states  $f_j$ ,  $j \in \{-n_f, \dots, n_f\}$ , i.e.,

$$\bar{p}_f = \sum_{j=-n_f}^{n_f} \tilde{p}_{f_j}. \quad (59)$$

Let

$$\alpha_f \triangleq \sum_{j=-n_f}^{n_f} \eta_{f_j} \cdot \exp(\beta \frac{j}{n_f} \Delta_f), \quad \forall f \in \mathcal{F}. \quad (60)$$

Then we have

$$\bar{p}_f = \frac{\alpha_f \exp(\beta x_f)}{\sum_{f' \in \mathcal{F}} \alpha_{f'} \exp(\beta x_{f'})}, \quad \forall f \in \mathcal{F}. \quad (61)$$

By (27), we know

$$p_f^* = \frac{\exp(\beta x_f)}{\sum_{f' \in \mathcal{F}} \exp(\beta x_{f'})}, \quad \forall f \in \mathcal{F}. \quad (62)$$

Let

$$\bar{\alpha} \triangleq \frac{\sum_{f' \in \mathcal{F}} \alpha_{f'} \exp(\beta x_{f'})}{\sum_{f' \in \mathcal{F}} \exp(\beta x_{f'})}. \quad (63)$$

It is not hard to see that  $\frac{p_f^*}{\bar{p}_f} = \frac{\bar{\alpha}}{\alpha_f}$ , so

$$p_f^* \geq \bar{p}_f \text{ iff } \alpha_f \leq \bar{\alpha}. \quad (64)$$

The total variation distance

$$d_{TV}(\mathbf{p}^*, \bar{\mathbf{p}}) = \frac{1}{2} \sum_{f \in \mathcal{F}} |p_f^* - \bar{p}_f| \quad (65)$$

$$= \sum_{f \in A} (p_f^* - \bar{p}_f), \quad (66)$$

where  $A \triangleq \{f \in \mathcal{F} : p_f^* \geq \bar{p}_f\}$ .

By (64), we know  $A = \{f \in \mathcal{F} : \alpha_f \leq \bar{\alpha}\} \subset \mathcal{F}$ .

Therefore,  $\forall f \in A$ ,

$$p_f^* - \bar{p}_f = \frac{\exp(\beta x_f)}{\sum_{f' \in \mathcal{F}} \exp(\beta x_{f'})} - \frac{\alpha_f \exp(\beta x_f)}{\sum_{f' \in \mathcal{F}} \alpha_{f'} \exp(\beta x_{f'})} \quad (67)$$

$$= \frac{\exp(\beta x_f)}{\sum_{f' \in \mathcal{F}} \exp(\beta x_{f'})} - \frac{\alpha_f \exp(\beta x_f)}{\bar{\alpha} \sum_{f' \in \mathcal{F}} \exp(\beta x_{f'})} \quad (68)$$

$$= \frac{\exp(\beta x_f)}{\sum_{f' \in \mathcal{F}} \exp(\beta x_{f'})} \left[1 - \frac{\alpha_f}{\bar{\alpha}}\right]. \quad (69)$$

Since  $\sum_{j=-n_f}^{n_f} \eta_{fj} = 1$  and  $\forall j \in \{-n_f, \dots, n_f\}$ ,

$$\exp(\beta \frac{j}{n_f} \Delta_f) \geq \exp(-\beta \Delta_f) \geq \exp(-\beta \Delta_{\max}) \quad (70)$$

and

$$\exp(\beta \frac{j}{n_f} \Delta_f) \leq \exp(\beta \Delta_f) \leq \exp(\beta \Delta_{\max}), \quad (71)$$

by (60) we know that  $\forall f \in \mathcal{F}$

$$\alpha_f \geq \sum_{j=-n_f}^{n_f} \eta_{fj} \cdot \exp(-\beta \Delta_{\max}) = \exp(-\beta \Delta_{\max}) \quad (72)$$

and

$$\alpha_f \leq \sum_{j=-n_f}^{n_f} \eta_{fj} \cdot \exp(\beta \Delta_{\max}) = \exp(\beta \Delta_{\max}). \quad (73)$$

Then by (63), we have  $\bar{\alpha} \leq \exp(\beta \Delta_{\max})$ . Therefore,

$$1 - \frac{\alpha_f}{\bar{\alpha}} \leq 1 - \frac{\exp(-\beta \Delta_{\max})}{\exp(\beta \Delta_{\max})} = 1 - \exp(-2\beta \Delta_{\max}), \forall f \in A \subset \mathcal{F}. \quad (74)$$

So by (69), we have  $\forall f \in A$ ,

$$p_f^* - \bar{p}_f = \frac{\exp(\beta x_f)}{\sum_{f' \in \mathcal{F}} \exp(\beta x_{f'})} \left[1 - \frac{\alpha_f}{\bar{\alpha}}\right] \quad (75)$$

$$\leq \frac{\exp(\beta x_f)}{\sum_{f' \in \mathcal{F}} \exp(\beta x_{f'})} (1 - \exp(-2\beta \Delta_{\max})). \quad (76)$$

Then,

$$d_{TV}(\mathbf{p}^*, \bar{\mathbf{p}}) = \sum_{f \in A} (p_f^* - \bar{p}_f) \quad (77)$$

$$\leq \sum_{f \in A} \frac{\exp(\beta x_f)}{\sum_{f' \in \mathcal{F}} \exp(\beta x_{f'})} (1 - \exp(-2\beta \Delta_{\max})) \quad (78)$$

$$\leq \sum_{f \in \mathcal{F}} \frac{\exp(\beta x_f)}{\sum_{f' \in \mathcal{F}} \exp(\beta x_{f'})} (1 - \exp(-2\beta \Delta_{\max})) \quad (79)$$

$$= 1 - \exp(-2\beta \Delta_{\max}). \quad (80)$$

Therefore,

$$|\mathbf{p}^* \mathbf{x}^T - \bar{\mathbf{p}} \mathbf{x}^T| = \left| \sum_{f \in \mathcal{F}} (p_f^* - \bar{p}_f) x_f \right| \quad (81)$$

$$\leq x_{\max} \sum_{f \in \mathcal{F}} |p_f^* - \bar{p}_f| \quad (82)$$

$$= 2x_{\max} d_{TV}(\mathbf{p}^*, \bar{\mathbf{p}}) \quad (83)$$

$$\leq 2x_{\max} (1 - \exp(-2\beta \Delta_{\max})). \quad (84)$$

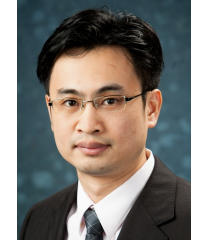
This concludes the proof.



**Shaoquan Zhang** received his B.Eng. degree from the University of Science and Technology of China in 2008. Currently he is a PhD student at The Chinese University of Hong Kong. His research interests include P2P system analysis and algorithm design, distributed and stochastic network optimization.

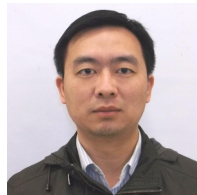


**Ziyu Shao** received the B.S. and M.S. degrees in electrical engineering from Peking University, Beijing, China, in 2001 and 2004, respectively and the Ph.D. degree in information engineering from The Chinese University of Hong Kong, Shatin, N.T., Hong Kong, in 2010. Currently, he is a Postdoctoral Fellow at The Chinese University of Hong Kong. His research interests include network coding, information theory, cloud computing, mobile computing, network optimization and distributed algorithm design.



**Minghua Chen** received his B.Eng. and M.S. degrees from the Dept. of Electronic Engineering at Tsinghua University in 1999 and 2001, respectively. He received his Ph.D. degree from the Dept. of Electrical Engineering and Computer Sciences at University of California at Berkeley in 2006. He spent one year visiting Microsoft Research Redmond as a Postdoc Researcher. He joined the Dept. of Information Engineering, the Chinese University of Hong Kong in 2007 as an Assistant Professor. He is also an Adjunct Associate Professor in Peking

University Shenzhen Graduate School in 2011-2014. He received the Eli Jury Doctoral Thesis award from UC Berkeley in 2007, the ICME Best Paper Award in 2009, the IEEE Trans. on Multimedia Best Paper Award in 2009, and the ACM Multimedia Best Paper Award in 2012. His current research interests include smart (micro) grids, energy-efficient data centers, distributed and stochastic network optimization and control, multimedia networking, p2p networking, wireless networking, multi-level trust data privacy, network coding, and secure network communications.



**Libin Jiang** received his B.Eng. degree from the University of Science and Technology of China in 2003, M.Phil. degree from the Chinese University of Hong Kong in 2005, and Ph.D. degree in electrical engineering and computer sciences from the University of California at Berkeley in 2009. His research interests include networking, communications, game theory, and Smart Grid. Dr. Jiang received the David Sakris Memorial Prize for outstanding doctoral research, and the best presentation award in the ACM Mobihoc'09 S3 Workshop. He is currently

with Qualcomm Research.