European Journal of Operational Research 255 (2016) 333–344

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Discrete Optimization

Allocation of flows in closed bipartite queueing networks☆

James D. Brooks[a,1], Koushik Kar[b,*], David J. Mendonça[c]

[a] G.E. Global Research Center, Niskayuna, NY 12309, USA
[b] Department of Electrical, Systems, and Computer Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180, USA
[c] Department of Industrial and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180 USA

ARTICLE INFO

Article history:
Received 4 April 2015
Accepted 10 May 2016
Available online 27 May 2016

Keywords:
(I) Networks
(O) OR in disaster relief
Queueing systems
Routing control

ABSTRACT

This paper describes a novel method for allocating agents to routes in a closed bipartite queueing network to maximize system throughput using three open network approximations. Results are presented which compare this method with known prior work and optimal solutions to provide an empirical optimality gap. Average empirical optimality gaps of 1.29 percent, 1.13 percent and 1.29 percent are observed for the three approximations considered. Further, because many systems are under the control of rational agents, conditions are derived in order to determine properties of the market context that induce optimal behavior. It is shown that uniform rewards do not yield an efficient rational equilibrium in general. However, for systems with homogeneous servers and travel times or those with travel times that are much larger than queue waiting times, uniform rewarding is optimal.

1. Introduction

Optimization of queueing systems has a long history (Stidham, 2009). Kleinrock provides the following general categories of optimization of queueing systems: capacity assignment, flow assignment, and topology (Kleinrock, 1993). Bell and Stidham further distinguish between dynamic (control) strategies and design (or static) strategies (Bell & Stidham, 1983). Dynamic strategies are functions of the current system state, while design strategies are state-independent and put in place prior to the operation of the system. For this paper, static flow assignment (i.e., routing without knowledge of current system state) is the primary focus, particularly as it relates to designing a market context for systems under the control of rational dispatchers. Throughout this paper (as well as in the majority of the relevant literature), optimization refers generally to maximizing the throughput of a system.

The motivating application for this work is the removal of debris from a region affected by a disaster event. This debris removal mission involves a set of hauling vehicles that cyclically deliver debris from a potentially large number of pickup sites (typically at curbsides) to a much fewer number of temporary debris storage and reduction (TDSR) sites, forming a bipartite network. Improvement in the throughput of this system (i.e., the average number of loads delivered per hour) translates directly to the ability of a community to return to normal operation after the disaster event. In order to accomplish this task, contractors provide teams of people and equipment to load and haul debris from many pickup sites to a few central processing sites. Each of these sites operates as a first-come-first-served queue and the number of vehicles is approximately fixed. There is no method presently in place for regulating or prescribing the allocation of the vehicles circulating in the different routes (cycles) of the system. Rather, they are assigned by a contractor in such a way as to maximize their profit.

Due to the geographic extent of the system, a routing (allocation) policy which does not require full knowledge of the location of all agents is desired (i.e. a state-independent policy). While some limited prior work has considered optimal state-independent routing probabilities in closed bipartite queueing networks, no work is known to provide a method of partitioning the agents into *groups*, each with its own deterministic route, as is known to be optimal (Tripathi & Woodside, 1988). This work first presents a novel method for generating optimal routing probabilities in these networks using reasonable modeling approximations. It then presents a method for determining an entity partition, each with its own deterministic routing, from these probabilities. Methods for dealing with some fixed (given) probabilistic routing in the system are also discussed.

☆ This is an extended version of a paper that appeared in the *2013 IEEE International Conference on Technologies for Homeland Security*, Waltham, MA in November 2013, with the title “Dynamic allocation of entities in closed queueing networks: An application to debris removal”.

* Corresponding author. Tel.: +1 5182762653.
E-mail addresses: brooksja@ge.com (J.D. Brooks), kark@rpi.edu (K. Kar), mendod@rpi.edu (D.J. Mendonça).

[1] This work was done when he was affiliated with the Department of Industrial and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY, 12180 USA

http://dx.doi.org/10.1016/j.ejor.2016.05.017
0377-2217/

The next section (Section 2) describes related prior work in network routing and motivates the approach used. The formulation for the general allocation problem in closed queueing networks is then discussed in detail (Section 3), using a small example to illustrate the derivation before presenting the general result. The partition algorithm is then described before numerical results are presented. These results compare the new method with both existing methods and provably optimal allocations to obtain an estimate of the empirical optimality gap. Finally, numerical results are presented and discussed (Section 4), followed by some conclusions and directions for future work (Section 5).

2. Literature review

Prior research in flow routing of agents in general queueing networks spans a variety of fields, and may be classified according to the network topology of interest: parallel server system, open network, or closed network. Both centralized and decentralized routing strategies have been developed. Centralized approaches assume that the behavior of agents in the network can be fully controlled, while decentralized approaches allow that agents may be autonomous. The latter leads directly into control via market influence or other decentralized means. Both types of approaches are discussed in the remainder of this section.

2.1. Centralized routing

Prior work in centralized routing has largely focused on the *parallel server* topology, in which arriving external agents choose from among a number of parallel servers and then, once serviced, exit the system. The objective is to minimize the expected delay (or equivalently the expected total number of agents in the system). Both dynamic (control) and static (design) policies have been explored. The join-shortest-queue (JSQ) policy is optimal for identical exponential servers (Winston, 1977) and asymptotically optimal for the multi-server case (Houck, 1987). Other policies explored in this system topology include minimum expected delay, an extension of JSQ for heterogeneous servers (Lui, Muntz, & Towsley, 1995), and round robin (Liu & Towsley, 1994). The throughput benefit of dynamic routing (allocation) in closed networks has been shown to decrease with increasing number of agents in the system (Delasay, Kolfal, & Ingolfsson, 2012).

Broadly speaking, faster servers are preferred in the optimal routing policy. If the parallel system has only two heterogeneous servers, it is optimal to send all traffic to the faster server until the number of agents waiting exceeds a threshold, at which point the slower servers should be utilized (Filipiak, 1984; Lin & Kumar, 1984). Under heavy traffic, these state-dependent threshold policies converge to a static threshold policy (Teh & Ward, 2002). For multiple parallel servers, faster servers are, disproportionally, more highly utilized (Delasay et al., 2012; Piepmeier, 1975). In fact, when multiple servers are available to an incoming agent, the faster-server-first policy is known to be asymptotically optimal while allowing slower servers to idle (Armony, 2005).

Flow optimization in *open networks* has been studied using a nonlinear programming framework (Whittle, 1984) (Kameda & Zhang, 1995), a Markov decision process framework (Seidmann, 1988), and a Brownian heavy-traffic approximation (Kelly & Laws, 1993). A two-stage approach to optimizing multicommodity flows in generalized networks using network approximations was developed in Morabito, de Souza, and Vazquez (2014). Whittle describes a method of maximizing the saturation throughput of an open network by selecting routing probabilities (Whittle, 1984). Optimal flows are not unique in general, though the relative server utilizations are so Kameda and Zhang (1995). Combining dynamic routing and sequencing of a network comprising of multiple agent classes (each with their own dedicated queues at each server (Kelly & Laws, 1993)), a pathwise solution outperforms optimal stochastic routing.

Throughput-maximizing allocation in *closed networks* has been explored (Kobayashi & Gerla, 1983) using a steepest-ascent method, in which the throughput of the system is given by the ratio of the network normalizing constant for $N-1$ and N agents. The method relies on the fact that the throughput function is (pseudo)concave in workload. The result is an optimal probabilistic routing matrix. These results have recently been extended to the multi-server case (Smith, 2011). A major drawback of these iterative methods (using Mean Value Analysis, or MVA) is their inability to produce optimality conditions which can meaningfully inform policy decisions.

To the best of our knowledge, there is no prior work that describes methods of partitioning a fixed population of agents to deterministic routes in a closed queueing network in order to maximize the system throughput. This is somewhat surprising given that the optimal routing strategy in product-form networks have been characterized (Cheng & Muntz, 1991; Hordijk & Loeve, 2000; Tripathi & Woodside, 1988).

2.2. Decentralized routing

Much research on queueing systems assumes that the behavior of agents flowing through the system is deterministic or nearly so. These agents operate with a predefined utility function and assumed deterministic behavior based on the value of this function. The relevant bodies of literature are rational behavior in queueing systems (Hassin & Haviv, 2003) and network congestion games (Roughgarden, 2005).

Wardrop equilibrium is an equilibrium notion often used in the context of network routing and traffic flow (Wardrop, 1952). A general network is composed with delay functions on each arc which are a function of the flow (traffic) on the arc. This type of equilibrium has been explored with linear, nonlinear, and queueing delay functions (Anselmi, Ayesta, & Wierman, 2011; Roughgarden, 2003). A network is said to be in this kind of equilibrium when all possible routes in use for each origin-destination pair have equal delay (i.e., no individual can benefit by unilaterally changing between routes in use). Furthermore, for routes between the same origin-destination pair which have no traffic (i.e., those that are unused for this origin-destination pair in equilibrium), the delay for a single vehicle must be greater than the delay of the routes in use. Finding such an equilibrium can be non-trivial. Indeed, an entire body of literature has developed around this problem (the traffic assignment problem (Correa & Stier-Moses, 2010)). Wardrop equilibrium is an instance of Nash equilibrium where the number of agents is infinitely large (i.e., in *non-atomic* games) (Correa & Stier-Moses, 2010). It is also less restrictive in the kinds of cost functions allowed. For this work, this equilibrium concept is extended to closed networks which have parallel cyclic paths:

Definition 1. A closed network is said to be in Wardrop-like equilibrium when all used cycles have the same mean loop delay and this common delay exceeds that of any unused cycles.

It is expected that for cases where a uniform payment policy is in effect, an economic, rational dispatcher will try to attain a Wardrop-like equilibrium. However, the payment policy is not uniform in practice. Because these prior definitions assume that the net profit gained by making the trip along any path (or cycle) is identical, the more complete notion of *economic Wardrop-like equilibrium* is introduced.

Definition 2. A closed network is said to be in economic Wardrop-like equilibrium when all used cycles have the same expected