Computers & Industrial Engineering 99 (2016) 260–268

Contents lists available at ScienceDirect

Computers & Industrial Engineering

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

Optimal network flow: A predictive analytics perspective on the fixed-charge network flow problem

Charles D. Nicholson, Weili Zhang *

School of Industrial and Systems Engineering, University of Oklahoma, USA

ARTICLE INFO

Article history:
Received 2 October 2015
Received in revised form 4 July 2016
Accepted 29 July 2016
Available online 1 August 2016

Keywords:
Network analysis
Fixed charge network flow
Predictive modeling
Critical components

ABSTRACT

The fixed charge network flow (FCNF) problem is a classical NP-hard combinatorial problem with wide spread applications. To the best of our knowledge, this is the first paper that employs a statistical learning technique to analyze and quantify the effect of various network characteristics relating to the optimal solution of the FCNF problem. In particular, we create a probabilistic classifier based on 18 network related variables to produce a quantitative measure that an arc in the network will have a non-zero flow in an optimal solution. The predictive model achieves 85% cross-validated accuracy. An application employing the predictive model is presented from the perspective of identifying critical network components based on the likelihood of an arc being used in an optimal solution.

1. Introduction

The fixed charge network flow problem (FCNF) can be easily described as follows. For a given network, each node may have a supply or demand commodity requirement and each incident arc have variable and/or fixed costs associated with commodity flow. The aim of the FCNF is to select the arcs and assign feasible flow to them in order to transfer commodities from supply nodes to demand nodes at a minimal total cost. The transportation problem (Balinski, 1961; El-Sherbiny & Alhamali, 2013), lot sizing problem (Steinberg & Napier, 1980), facility location problem (Aikens, 1985; Daskin, 1995), network design problem (Costa, 2005; Ghamlouche, Crainic, & Gendreau, 2003; Lederer & Nambimadom, 1998) and others (Armacost, Barnhart, & Ware, 2002; Jarvis, Rardin, Unger, Moore, & Schimpeler, 1978) can be modeled as a FCNF.

The FCNF problem is known to be NP-hard (Guisewite & Pardalos, 1990). A significant amount of effort has been invested to study and develop efficient approaches to the FCNF. Many techniques commonly utilize branch and bound to search for an exact solution to the FCNF (Barr, Glover, & Klingman, 1981; Cabot & Erenguc, 1984; Driebeek, 1966; Hewitt, Nemhauser, & Savelsbergh, 2010; Kennington & Unger, 1976; Ortega & Wolsey, 2003; Palekar, Karwan, & Zionts, 1990). Branch and bound however may be inefficient due to lacking tight bounds during the linear relaxation step. Heuristic approaches to find the near-optimal solution of the FCNF have generated considerable research interest (Adlakha & Kowalski, 2010; Antony Arokia Durai Raj, 2012; Balinski, 1961; Kim & Pardalos, 1999; Molla-Alizadeh-Zavardehi, Hajiaghaei-Keshteli, & Tavakkoli-Moghaddam, 2011; Monteiro, Fontes, & Fontes, 2011; Sun, Aronson, McKeown, & Drinka, 1998). State-of-the-art MIP solvers combine a variety of cutting plane techniques, heuristics and the branch and bound algorithm to find the global optimal solution. Modern MIP solvers use preprocessing methods to reduce the search space by taking information from the original formulations, which significantly accelerate the solving processes (Bixby, Fenelon, Gu, Rothberg, & Wunderling, 2000). In this paper, we take a decidedly different approach to leveraging information from the problem formulation and FCNF instances. That is, we are interested in gaining information about how the various topological and component characteristics relate to the selection of arcs used to transmit the optimal flow. At this time, we have found no literature that approaches a study of the FCNF problem from the perspective of statistical learning.

FCNF formulations are useful in many practical problems. Modern societies are heavily dependent on distributed systems, e.g. communication networks (Cohen, Erez, Ben-Avraham, & Havlin, 2000), electric power transmission networks (Dobson, Carreras, Lynch, & Newman, 2007), and transportation networks (Zheng, Gao, & Zhao, 2007). Designing and maintaining such systems is an important research area in network science. In particular, developing resilient network infrastructures (i.e., resilient with respect to natural disasters or intentional attacks) is of utmost importance and the ability to identify critical components in complex networks

* Corresponding author at: School of Industrial and Systems Engineering, University of Oklahoma, 202 W. Boyd St., Room 442, Norman, OK, USA.

E-mail address: weili.zhang-1@ou.edu (W. Zhang).

http://dx.doi.org/10.1016/j.cie.2016.07.030
0360-8352/

has reached a level of national urgency (Birchmeier, 2007). The destruction or damage of one or more critical components in a networked system could have significant consequences in terms of overall system performance (Bell, 2000; Smith, Qin, & Venkatanarayana, 2003). The definition of component criticality is often associated with an overall network performance metric. A component whose hypothetical failure most impacts the network performance level is identified as critical. A substantial body of work using a variety of methods has focused on identifying critical components within networks, e.g. topological approach (Bompard, Napoli, & Xue, 2009; Crucitti, Latora, & Marchiori, 2005), simulation (Eusgeld, Kröger, Sansavini, Schläpfer, & Zio, 2009), optimization (Bier, Gratz, Haphuriwat, Magua, & Wierzbicki, 2007; Shen, Smith, & Goli, 2012; Zio, Golea, & Rocco, 2012), service measure (Dheenadayalu, Wolshon, & Wilmot, 2004; Scott, Novak, Aultman-Hall, & Guo, 2006) and graph theory (Demšar, Špatenková, & Virrantaus, 2008). In this study we consider an application of our statistical model with respect to identifying critical components wherein the minimum total commodity routing cost, inclusive of fixed costs, is the overall network performance metric.

To the best of our knowledge no existing work has developed models to help characterize predictive network features of optimal solutions to the FCNF. More broadly, little work has been published so far in the application of statistical learning to traditional optimization or network problems. Rocco and Muselli (2004, 2005) developed a decision tree and a hamming clustering model to predict network connectivity reliability in graphs. Hamming clustering is applicable only if both the predicted value and all predictors are binary (Muselli & Liberati, 2002). The binary predictions relating to connectivity were made based on a single type of predictor – the status of each arc in the graph as either failed or operating. Based on this information they attempted to evaluate the reliability of origin-destination connectedness. Empirically they create one network instance (11 nodes, 21 edges) and randomly sample from the possible state space of edge failures. Among the possible 2^{21} states, 2000 were assigned to a training set and 1000 assigned to a test set. The models were developed on the 2000 training observations and highly accurate predictions were observed on the test set. While the predictive models developed were highly accurate, they are inherently linked to the single network instance considered.

In this study we employ a statistical learning technique to analyze the data associated with optimal FCNF solutions and we develop a relatively generalizable model based on several salient network features to predict which arcs will be used in an optimal solution. By solving thousands of generated FCNF instances we collect over 60,000 observations and develop a logistic regression model based on the dataset. This model allows us to quantify the influence of several important network characteristics. The resulting model has several potential applications. In this study, we demonstrate an application for providing an alternative approach to identifying critical network components. The remainder of this paper is organized as follows. Section 2 introduces the background of the FCNF and the logistic regression model. The process for developing the predictive model is discussed in Section 3. The identification of critical components using the model is presented in Section 4. Section 5 summarizes the results and introduces planned future work.

2. Background

2.1. Fixed charge network flow problem

The fixed charge network flow (FCNF) problem is described on a network $G = (N, A)$, where N and A are the set of nodes and arcs, respectively. Let c_{ij} and f_{ij} denote the variable and fixed cost of arc $(i,j) \in A$, respectively. Each node $i \in N$ has a commodity requirement r_i associated with it (if it is a supply node, $r_i > 0$; if a demand node, $r_i < 0$; if a transshipment node, $r_i = 0$). An arc parameter M_{ij} is used in the problem formulation to ensure that the fixed cost f_{ij} is incurred whenever there is a positive flow on arc $(i,j) \in A$. There are two decision variable types: y_{ij} which denotes the decision variable to use arc $(i,j) \in A$ in a solution and x_{ij} denotes the commodity flow on (i,j). The mathematical formulation is as follows,

$$\min \sum_{(i,j)\in A} (c_{ij}x_{ij} + f_{ij}y_{ij}) \tag{1}$$

$$\text{s.t.} \sum_{(i,j)\in A} x_{ij} - \sum_{(j,i)\in A} x_{ji} = r_i \quad \forall i \in N \tag{2}$$

$$0 \leqslant x_{ij} \leqslant M_{ij}y_{ij} \quad \forall (i,j) \in A \tag{3}$$

$$y_{ij} \in \{0,1\} \quad \forall (i,j) \in A \tag{4}$$

Constraint (2) ensures that the inflow and outflow satisfy the supply/demand at node $i \in N$. The parameter M_{ij} in constraint (3) is either the associated arc flow capacity or an artificial arc capacity (for uncapacitated problems). The constraint ensures that the flow on arc $(i,j) \in A$ can be positive only when the arc $(i,j) \in A$ is open ($y_{ij} = 1$). If arc (i,j) does not have a capacity, M_{ij} should be set to a value which is large enough to not inhibit the optimal flow. All problems in this study are uncapacitated and each M_{ij} is set to the total supply in the network. Constraint (4) defines y_{ij} as binary, which makes the problem a 0–1 mixed integer programming problem.

2.2. Logistic regression

Logistic regression is a widely-used technique for classification modeling and is commonly used in business modeling, data mining applications, biological fields, and others (Camdeviren, Yazici, Akkus, Bugdayci, & Sungur, 2007; Hosmer, Lemeshow, & Sturdivant, 2013; Menard, 2002). While there are many classification modeling techniques (e.g., support vector machines, random forests, boosted trees), logistic regression has an advantage regarding model interpretability. Decision trees which are also easy to interpret have a drawback in that they are often unstable. That is, the rules generated by a decision tree are highly sensitive to the instance of training data (Friedman, Hastie, & Tibshirani, 2001). Given that we are interested in analyzing data to understand the characteristics of optimal FCNF solutions, interpretability and stability are important.

We denote the dependent variable (also called a response variable) as Y and define it as follows,

$$Y = \begin{cases} 1, & \text{arc has positive flow in the FCNF optimal solution} \\ 0, & \text{otherwise.} \end{cases}$$

The logistic regression function produces a probability that the response variable equals 1 given the data values observed for the associated k predictor variables, $p_1, \ldots, p_k$,

$$P(Y = 1 | p_1, \ldots, p_k) = \frac{1}{1 + e^{-\left(\beta_0 + \sum_{i=1}^{k} \beta_i p_i\right)}} \tag{5}$$

where the parameters $\beta_1, \ldots, \beta_k$ are regression coefficients determined using maximum likelihood estimation during the modeling process. For a given set of observed values, the binary response variable is set to 1 if the predicted probability exceeds a cut-off point. The details for setting the cut-off value are discussed in Section 3.4.

The β values in a logistic regression model are interpreted similar to linear regression, in that, they represent partial slopes