QRF: an Optimization-Based Framework for Evaluating Complex Stochastic Networks – Online Supplement

Giuliano Casale, Imperial College London, UK, g.casale@imperial.ac.uk
Vittoria De Nitto Personé, University of Rome Tor Vergata, Italy, denitto@info.uniroma2.it
Evgenia Smirni, College of William and Mary, VA, US, esmirni@cs.wm.edu

Categories and Subject Descriptors: C.4 [**Performance of Systems**]: Modeling Techniques

General Terms: Algorithms, Performance

Additional Key Words and Phrases: Queueing, blocking, temporal dependence, state-dependence

Multiple Finite Capacity Queues

As observed in the paper, for simplicity we have described the QRF analysis in the case of a single finite capacity queue. The case with two or more finite capacity queues follows the same arguments but introduces complications due to possible chains of dependence, as shown in the following example.

Consider a station i blocked by a finite capacity queue f_1 such that $Head(\boldsymbol{m}_{f_1}) = i$. Suppose now that f_1 is itself blocked by a finite capacity queue f_2 such that $Head(\boldsymbol{m}_{f_2}) = f_1$ and assume that f_2 can route jobs to i. Then the marginals in (8) require to express the event in which f_2 sends a job to i and, due to the chain of dependencies, simultaneously f_1 sends a job to f_2 and i sends a job to f_1. This event does not contribute to the marginal balance in (8) because the population at n_i remains unchanged.

We here note that this problem does not arise if finite capacity stations are not blocked by each other. This case can then be handled by either avoiding a direct connection between finite capacity stations, or by introducing a surrogate infinite server station, with a large service rate, between f and g. The last option avoids the case where two unblocking events are perfectly simultaneous. This comes at the cost of adding an extra station for each pair of finite capacity queues, but it sufficient to address the problem and generalize the approach in an approximate manner.

Bounding Problem 1

We study a network composed of $M = 3$ queues, with queue $f = 1$ having a finite capacity of $F_1 = 4$ jobs and we use populations $N \geq F_1 + 1$ such that blocking can occur. We assume that service times are exponentially distributed at queues 2 and 3, while the finite capacity queue 1 has temporal dependent MAP service. For ease of interpretation, we consider different routing matrices resulting in different levels of balancing between *service demands*, which are mean number of visits divided by mean service rates. It is well-known that service demands, rather than just visits or rates, determine the effective utilization levels of resources in a system [Lazowska et al. 1984]. Specifically, we consider mean service rates $\mu_i = 1$, $i = 1, \ldots, M$, and the following routing matrix

$$P = \begin{bmatrix} 0.10 & 0.50 & 0.40 \\ p & 0 & 1-p \\ 0 & 0.50 & 0.50 \end{bmatrix} \quad (1)$$

which provides mean number of visits by computing the equilibrium of P interpreted as a discrete time Markov chain. We then vary the routing probability p to obtain three network profiles:

- $p = 0.99$*: maximum demand at finite capacity queue.* This is an unbalanced network where jobs cumulate at queue 1 the maximum amount of service time.
- $p = 0.81$*: maximum demand at infinite capacity queue.* This case is symmetric to $p = 0.99$ and queue 3 is now the bottleneck resource.
- $p = 0.90$*: balanced demands.* Setting $p = 0.90$ balances demands such that the cumulative time spent in service at the queues is identical.

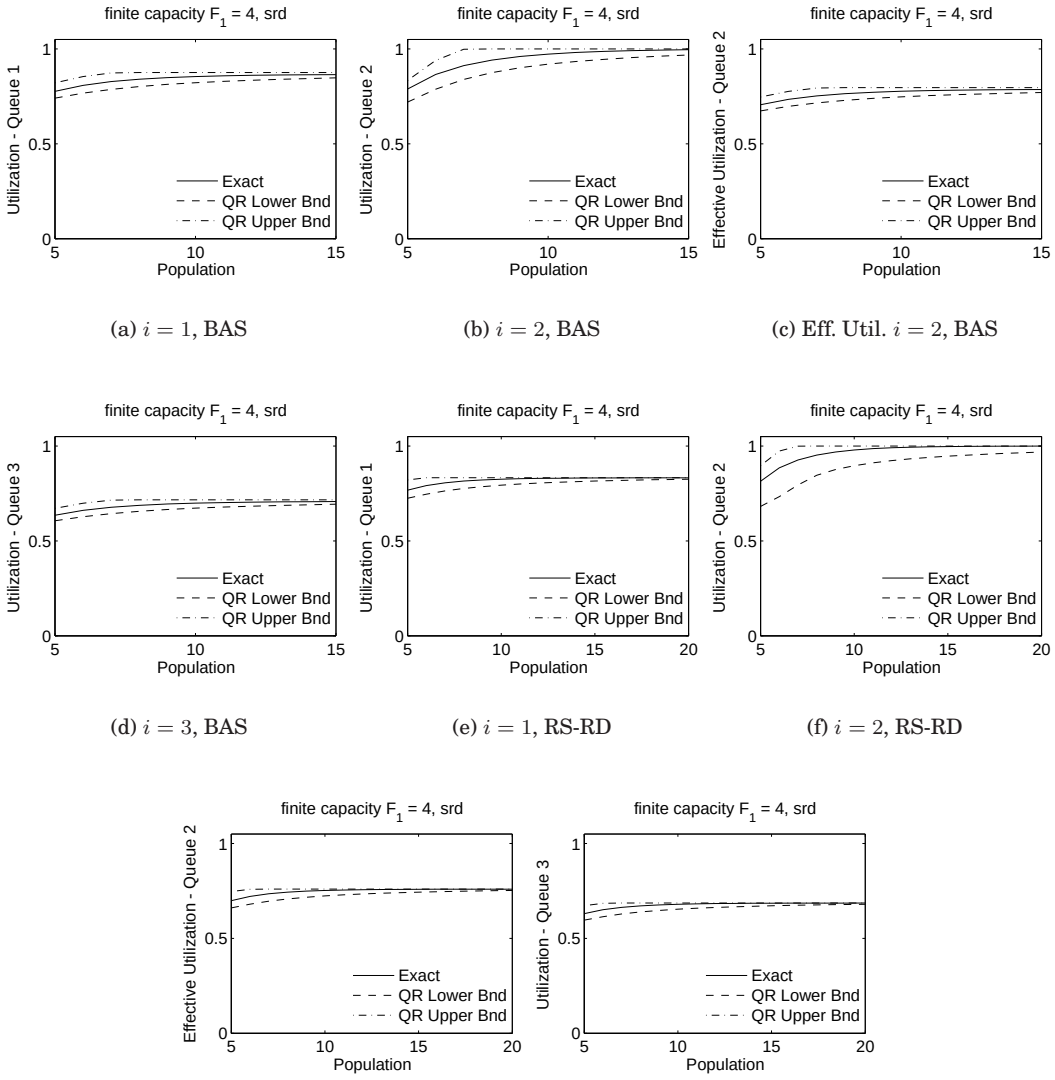
Figures 1-3 illustrate bounding results across more than 240 runs of BQR bounds and reports results for the utilization levels U_i and for the effective utilization E_2 of queue 2. Figures 1 reports BAS and RS-RD results, while the remaining figures only BAS results since RS-RD results are qualitatively very similar to the BAS ones. Note that it is $U_1 = E_1$ and $U_3 = E_3$ for all populations, thus we only report the effective utilization for queue 2. Since throughput and system response time follow easily from such quantities, such results are representative of the bounding quality for several performance metrics. The computational costs for a single run of BQR bounds are very small: on a laptop computer in the worst case the solution took 0.3s and 32MB for BAS and 0.2s and 7MB for RS-RD, with the difference being due to the small set of BQR probabilities in RS-RD due to $\boldsymbol{m} = \emptyset$.

The results indicate that the BQR bounds perform well in limiting the utilization and the effective utilization of the three queues. In several cases, the bounds are extremely tight with good results for the utilization being achieved in Figure 2 where queue 3 has the highest demand, while the effective utilization is tightly limited particularly in Figure 1. The results indicate that the upper and lower bounds perform equally well and that there are not substantial accuracy differences between BAS and RS-RD. The hardest quantity to bound is the utilization of queue 2, which is the sum of the effective utilization and the component due to the blocking of queue 2. Still, in such cases, the absolute gap between bound and exact value of the effective utilization is approximately 8%, whereas it is usually $3 - 4\%$ in the other queues. This makes the case that BQR bounds are effective for quantitative analysis of systems with blocking.

Bounding Problem 2: Comparison with EMVA Algorithm

In this section, we consider BAS blocking and compare the proposed method with an approximation proposed in the literature, namely the expanded mean value analysis (EMVA) algorithm proposed in [Yuzukirmizi 2006]. This algorithm considers marginal queue-length probabilities and uses them to estimate blocking probabilities alongside with effective service rates. The EMVA algorithm only supports exponential distributions, thus we illustrate the behaviour of the BQR bounds in this setting.

We have implemented the EMVA algorithm and validated it on a set of models. In several case, both EMVA and BQR bounds were very close to the optimal solution. In several other cases, we noticed that EMVA as the load increases may incur into instabilities that are not instead experienced by the BQR bounds. An example model is the case $p = 0.9$ in the previous set of examples. Results are shown in Figure 4 for a range of populations between 5 and 30 jobs. The figure shows the effective utilization at queue 2. The curves indicate that EMVA is able to follow well the trend of the exact solution for low populations. As the effective utilization grows larger towards its asymptotic value, EMVA becomes progressively less stable and the solution deteriorates in heavy load. Conversely, the BQR bounds maintain their bounding properties in a stable manner across the whole range of considered populations. This suggests that BQR bounds can be more reliable compared to existing techniques; however, the example also shows that there can be situations in which the bounds do not converge early, thus point approximations may not be accurate enough. We address these cases



(a) $i = 1$, BAS (b) $i = 2$, BAS (c) Eff. Util. $i = 2$, BAS

(d) $i = 3$, BAS (e) $i = 1$, RS-RD (f) $i = 2$, RS-RD

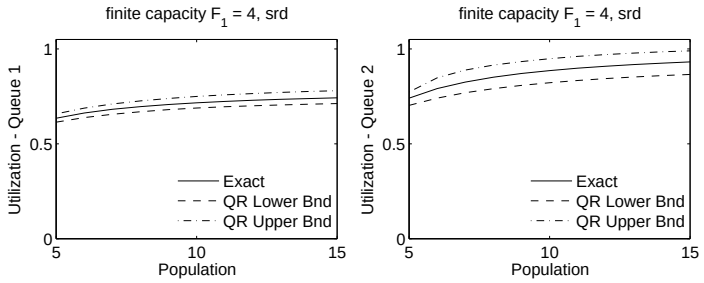
(g) Eff. Util. $i = 2$, RS-RD (h) $i = 3$, RS-RD

Fig. 1. Finite capacity queue 1 has highest demand ($p = 0.99$). MAP is short-range dependent (srd).

in Section 7 by developing two methodologies for approximation of queueing network models with blocking.

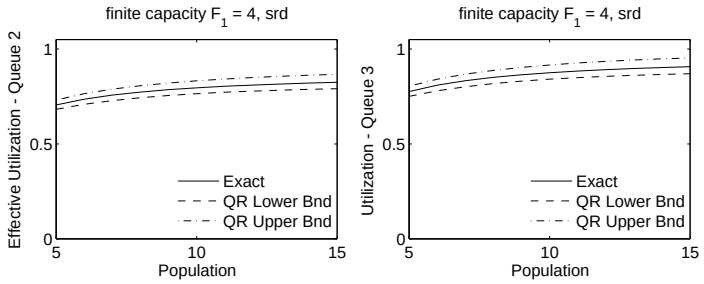
Approximation Problem 1: RS-RD Blocking

Let us first consider a model composed of $M = 5$ queues with $N = 10$ jobs, capacity $F_i = 5$ for each queue $i = 1, \ldots, M$, and service processes all equal to the short-range dependent MAP given in (12). Hence, all stations can be blocked. The routing matrix



(a) $i = 1$, BAS

(b) $i = 2$, BAS



(c) Eff. Util. $i = 2$, BAS

(d) $i = 3$, BAS

Fig. 2. Infinite capacity queue 3 has highest demand ($p = 0.81$). MAP is short-range dependent (srd).

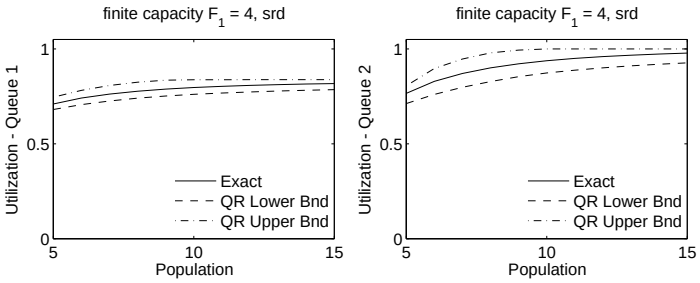
is

$$P = \begin{bmatrix} 0 & 0.5000 & 0 & 0 & 0.5000 \\ 0.5000 & 0 & 0.5000 & 0 & 0 \\ 0 & 0.5000 & 0 & 0.5000 & 0 \\ 0 & 0 & 0.5000 & 0 & 0.5000 \\ 0.5000 & 0 & 0 & 0.5000 & 0 \end{bmatrix}$$

This is a case where we compare approximations and bounds under multiple RS-RD blocking. We see in Figure 5 that the upper and lower bounds ("ub" and "lb", respectively) are not able to generate a tight envelope around the exact utilization and exact effective utilizations ("ex"). However, both MEM and MMI return almost perfect results within less than 2% utilization. Similarly to the toy example, MMI appears slightly more effective than MEM for capturing the probability distribution. Notice also that the MEM solution is slightly affected by numerical perturbations due to the fully symmetric routing of this network.

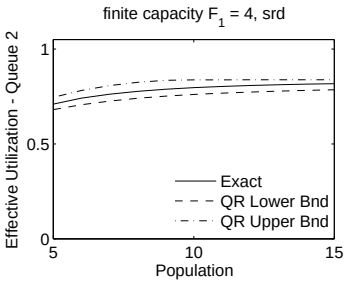
Approximation Problem 2: Central Server Model

We now consider a classic central-server-type topology, where queue 1 feeds parallel stations. The model is quite similar to the one used for the Bounding Problem in Section 6.1. We assume $M = 5$, $N = 10$, and routing matrix P^- as in Section 6.1. Similarly, service processes and finite capacities are identical to the ones in Section 6.1. Figure 6 reports experimental results. We see again that the proposed approximations are very effective, however this illustrates a case where also the bounds are very tight, and one

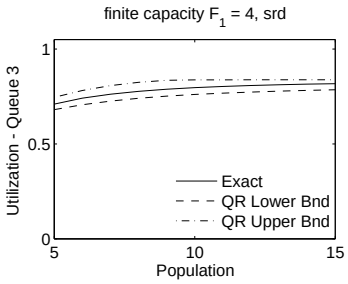


(a) $i = 1$, BAS

(b) $i = 2$, BAS



(c) Eff. Util. $i = 2$, BAS



(d) $i = 3$, BAS

Fig. 3. Balanced demands ($p = 0.90$). MAP is short-range dependent (srd).

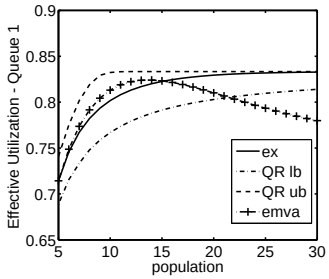
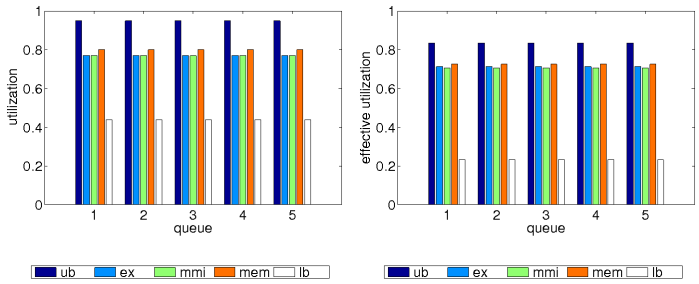


Fig. 4. Comparison with EMVA algorithm.

may for instance take their middle point as a first approximation of the exact value of the utilizations. Thus, this shows a case where station 1 has a dramatic difference between utilization and effective utilization, due to the blocking on queue 5. This is perfectly captured by our techniques.

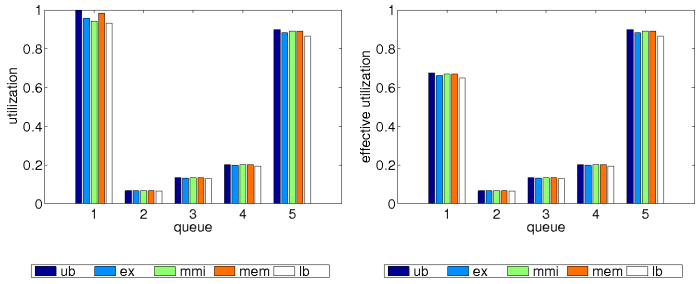
Application Domains

Queueing networks with blocking have been systematically investigated for decades and there are many examples of their applications to real-world systems. As mentioned in the introduction, models with blocking that can be analyzed with QRF are particularly suited for performance analysis of computer and communication systems,



(a) Utilization (b) Effective Utilization

Fig. 5. Approximation Problem 1 - queueing network model with RS-RD blocking



(a) Utilization (b) Effective Utilization

Fig. 6. Approximation Problem 2 - A queueing network model with central-server-type topology

albeit they have also been applied to other fields such as manufacturing and health care. The reader may consult surveys such as [Balsamo et al. 2001], [Onvural 1990], [Perros 1989], [Perros 1994] for specific examples of how blocking mechanisms such as RS-RD and BAS apply to real systems. Additional resources include papers focused on computer systems [Almeida and Kellert 2000], communication systems and networks [Awan et al. 2006], [Daduna and Holst 2008], streaming systems [Xia et al. 2007], manufacturing systems [Yamada et al. 2009], software architectures [Balsamo et al. 2003] and health care [Koizumi et al. 2005]. While in general RS-RD and BAS may not be exact models of the actual blocking happening in a real-world system, RS-RD is a reasonable modeling approximation for cases where a call to a finite capacity node does *not* block the caller. Instead, the BAS blocking represents the opposite case where the caller is blocked by this synchronization. This appears adequate to represent the blocking patterns of many enterprise workloads, as for instance illustrated in several examples in [Balsamo et al. 2003].

In the context of computer systems, since the QRF methodology is particularly suited for optimization, it could be easily integrated for optimal decision-making in problems such as load-balancing, optimal sizing of thread pools and buffers, and optimal resource allocation. The parallel system analyzed in Section 8 and the networks with state-dependent routing in Section 3.5 illustrate applicability to load-balancing. Another example is that of server consolidation problems in computer infrastruc-

tures [Ardagna et al. 2014]. These problems involve deciding the number of servers that will be utilized in an infrastructure to serve web requests, taking into account requirements on the utilization of the machines, end-to-end response times, and operational costs [Ardagna et al. 2014]. These problems are typically NP-hard and the goal is to find a good local optimum. The underpinning optimizations are mixed integer nonlinear programs where either the modelled simplifies the queueing analysis to basic M/M/1 and M/G/1 queues, or the decision variables are decomposed so that solves can rely on external procedures to evaluate a queueing subproblem at each iteration. Several algorithms exist to evaluate a queueing subproblem, e.g. AMVA [Bolch et al. 2006] and fluid queueing solvers [Franceschelli et al. 2013], however none of these methods offers robust approximations for dependent workloads, blocking, state-dependent routing, especially when considered in combination. QRF instead offers analysis methods that can encompass all of these features. Furthermore, it is well-suited for successive invocations, being able to quickly re-optimize the optimal solution of a complex queueing network using the methodology discussed in Section 8. It therefore offers a suitable alternative to existing external solvers for queueing subproblems.

In terms of expressiveness, compared to existing models with blocking, QRF allows to express temporal dependent service processes, which recent work has identified as being important for performance characterization of web servers [Mi et al. 2007] and disk drives [Riska and Riedel 2006]. In these systems, the service process is normally autocorrelated due to caching. However, it is difficult to express temporal dependent requirements in ordinary networks, particularly closed ones. Closed networks are important in computer system performance analysis to express limits on concurrency levels in accessing connections or gaining control of a thread, and finite capacity adds to the closed feature to express limits on buffer sizes that compound to the finite concurrency levels. Standard queueing network with blocking can be represented by closed systems and finite buffers, however only QRF models currently allow for temporal dependent descriptions of the service process.

REFERENCES

D. De Almeida and Patrick Kellert. 2000. Markovian and analytical models for multiple bus multiprocessor systems with memory blockings. *Journal of Systems Architecture* 46, 5 (2000), 455–477. DOI:10.1016/S1383-7621(99)00006-5

D. Ardagna, G. Casale, M. Ciavotta, J.F. Prez, and W. Wang. 2014. Quality-of-service in cloud computing: modeling techniques and their applications. *Journal of Internet Services and Applications* 5, 1 (2014). DOI:10.1186/s13174-014-0011-3

I. Awan, A. Yar, and M. E. Woodward. 2006. Analysis of Queuing Networks with Blocking under Active Queue Management Scheme. In *ICPADS*. IEEE Computer Society, 61–68. DOI:10.1109/ICPADS.2006.25

S. Balsamo, V. De Nitto Personé, and P. Inverardi. 2003. A review on queueing network models with finite capacity queues for software architectures performance prediction. *Perform. Eval* 51, 2/4 (2003), 269–288. DOI:10.1016/S0166-5316(02)00099-8

S. Balsamo, V. De Nitto Personé, and R. Onvural. 2001. *Analysis of queueing networks with blocking*. Kluwer Academic Publishers, Boston, MA.

G. Bolch, S. Greiner, H. de Meer, and K. S. Trivedi. 2006. *Queueing Networks and Markov Chains*. 2nd ed., John Wiley and Sons.

H. Daduna and M. Holst. 2008. Customer Oriented Performance Measures for Packet Transmission in a Ring Network with Blocking. In *Measuring, Modelling and Evaluation of Computer and Communication Systems (MMB), 2008 14th GI/ITG Conference*, Falko Bause and Peter Buchholz (Eds.). VDE Verlag, 223–236. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=5755055

D. Franceschelli, D. Ardagna, M. Ciavotta, and E. Di Nitto. 2013. SPACE4CLOUD: A Tool for System PerformAnce and Cost Evaluation of CLOUD Systems. In *Proceedings of the 2013 International Workshop on Multi-cloud Applications and Federated Clouds (MultiCloud '13)*. ACM, New York, NY, USA, 27–34. DOI:10.1145/2462326.2462333

N. Koizumi, E. Kuno, and T. E. Smith. 2005. Modeling Patient Flows Using a Queuing Network with Blocking. *Health Care Management Science* 8, 1 (2005), 49–60. DOI:10.1007/s10729-005-5216-3

E. D. Lazowska, J. Zahorjan, G. S. Graham, and K. C. Sevcik. 1984. *Quantitative System Performance*. Prentice-Hall.

N. Mi, Q. Zhang, A. Riska, E. Smirni, and E. Riedel. 2007. Performance impacts of autocorrelated flows in multi-tiered systems. *Perform. Eval* 64, 9-12 (2007), 1082–1101. DOI:10.1016/j.peva.2007.06.016

R. O. Onvural. 1990. Survey of Closed Queueing Networks with Blocking. *Comput. Surveys* 22, 2 (June 1990), 83–121. DOI:10.1145/78919.78920

H. G. Perros. 1989. A Bibliography of Papers on Queueing Networks with Finite Capacity Queues. *Perform. Eval* 10, 3 (1989), 255–260. DOI:10.1016/0166-5316(89)90015-1

H. G. Perros. 1994. *Queueing Networks with Blocking*. Oxford University Press, Inc., New York, NY, USA.

A. Riska and E. Riedel. 2006. Long-Range Dependence at the Disk Drive Level. In *QEST*. IEEE Computer Society, 41–50. DOI:10.1109/QEST.2006.27

C. H. Xia, Z. Liu, D. F. Towsley, and M. Lelarge. 2007. Scalability of fork/join queueing networks with blocking. In *Proceedings of the 2007 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems, SIGMETRICS 2007, San Diego, California, USA, June 12-16, 2007*. ACM, 133–144. DOI:10.1145/1269899.1254898

T. Yamada, N. Mizuhara, H. Yamamoto, and M. Matsui. 2009. A performance evaluation of disassembly systems with reverse blocking. *Computers & Industrial Engineering* 56, 3 (2009), 1113–1125. DOI:10.1016/j.cie.2008.09.029

M. Yuzukirmizi. 2006. Performance Evaluation of Closed Queueing Networks with Limited Capacities. *Turkish Journal of Engineering and Environmental Sciences* 30 (2006), 269–283. http://journals.tubitak.gov.tr/engineering/issues/muh-06-30-5/muh-30-5-1-0509-1.pdf