The price of anarchy in Tor latency reduction

Dieter Fiems

Department of Telecommunications and Information Processing, Ghent University, Belgium
Email: Dieter.Fiems@UGent.be

The expected cell latency for multi-hop M/D/1 queueing networks in Wardrop equilibrium is calculated. The load is balanced such that no user can benefit from unilaterally changing its routes. The results are applied to a typical snapshot of the Tor anonymity network, for which an efficient approach is devised to calculate the equilibrium routing strategy. For this snapshot, latency in the Wardrop equilibrium is numerically compared with the globally optimal latency and the latency of the original routing method of Tor.

Introduction: Tor is an anonymity service which routes user traffic through a three-hop network prior to accessing the internet [1, 2]. In the original Tor network, users construct a random path (a circuit in Tor terminology) by choosing nodes with probability proportional to their bandwidth. This choice often leads to considerable latency, which has motivated research in alternative circuit selection algorithms (see, e.g., [3]). Of particular interest is this letter’s contribution towards the work of Herbert et al. [4]. These authors use queueing theory to show that latency-optimal circuit selection is a convex optimisation problem and find that the expected cell latency can be reduced by a factor of 9 in a typical Tor usage scenario. Optimal routing, however, assumes that the traffic of every user is routed in accordance with the globally optimal routing probability vector. When this is the case, a single user can unilaterally improve its performance by choosing a route with less latency. Therefore, the optimal routing probability vector is not stable in the sense that individual users may gain from using a different routing vector, thereby disrupting the optimal routing solution. In this letter, we focus on routing strategies which are stable in the sense that users cannot benefit from changes. We adopt the analytic setting of [4]: (i) each node is a single server queue with Poisson arrivals and deterministic service times; (ii) the latency is dominated by one direction such that we can focus on a unidirectional network; (iii) cells are routed probabilistically; (iv) the network status varies slowly such that it can be assumed to be constant; and (v) the same node is allowed to be used multiple times in the same path. The stable routing probability vector then constitutes a so-called Wardrop equilibrium. In a Wardrop equilibrium, the latency on all the routes that are actually used are equal, and this latency is smaller than the latency which would be experienced by a single user on any unused route. In general, calculating the Wardrop equilibrium is a convex optimisation problem [5]. However, we show that the Wardrop equilibrium of the Tor network can be found by solving a limited number of scalar root finding problems.

Wardrop equilibrium: We consider a network with \( n \) nodes (node 1 to node \( n \)) and \( N \) hops (hop 1 to hop \( N \)) and \( \mu_i \) denote the service rate of the \( i \)-th node, \( i \in \{1, \ldots, n\} \). Cells are randomly routed through this network, requiring service by a single node at each hop. Let \( \lambda_i \) be the arrival rate of cells to be routed through the network. At the network, requiring service by a single node at each hop. Let \( \lambda^* \) denote the critical arrival rate which the network can carry, that is, there exist a stable policy for all \( \lambda < \lambda^* \). In the general case, there is no simple closed-form expression for \( \lambda^* \), but one needs to solve the following linear optimisation problem instead for \( \alpha = 1/\lambda^* \).

\[
\min_{ \alpha \geq 0 } \alpha ,
\]

\[
\text{subject to } \frac{1}{\mu_i} \sum_{j=1}^{N} \lambda_{ij} - \alpha \leq 0 , \quad \text{for } i = 1, \ldots, n ,
\]

\[
\sum_{i \in \Omega} \lambda_{ij} = 1 , \quad \text{for } j = 1, \ldots, N .
\]

The first constraint ensures that no queue has a load exceeding 1. The second constraint ensures that no traffic is lost between hops. When the nodes are grouped into distinct sets such that all the nodes within a set can be used at the same hop (or at the same set of hops), optimal routing (optimal) can be considerably simplified. We explore this approach for Tor in the next section, where we derive a simple expression for \( \lambda^* \).

Provided the nodes are not overloaded, and approximating the arrival processes at hops 2 to \( N \) by Poisson processes, the Pollaczek–Khinchine formula [5] gives the expected latency \( w_i \) for a packet at the \( i \)-th node, \( i = 1, \ldots, n \).

\[
w_i = \frac{1}{\mu_i} + \frac{\rho_i}{2 \mu_i (1 - \rho_i)} .
\]

As the \( \rho_i \) node is selected for the \( i \)-th hop with probability \( \rho_i \), we finally find the latency throughout the network,

\[
w = \sum_{i=1}^{n} \sum_{j=1}^{N} \rho_j \lambda_{ij} w_i = \sum_{i=1}^{N} \lambda_{ij} w_i / \lambda_i .
\]

Now consider any routing policy \( P \in \mathcal{P} \) and the corresponding latencies \( w_i \). If there are nodes with unequal expected latencies available at a hop, users can get a lower latency by adapting their routing policy to favour the best nodes. Hence, an admissible routing policy is only an (Wardrop) equilibrium policy provided that there exist \( P^* \) such that,

\[
w_i = w_i^* \quad \text{for } \rho_i > 0 ,
\]

\[w_i \geq w_i^* \quad \text{for } \rho_i = 0 \quad \text{and } i \in \Omega_i .
\]

The first equation states that all nodes that are used at a hop offer the same expected latency, while the second equation states that no nodes with a better latency can be used. The formulation above can be directly cast into the framework of Wardrop equilibria which allows one to assert the existence of such an equilibrium for \( \lambda < \lambda^* \) [6]. Solving the Wardrop equilibrium yields the arrival rates \( \lambda_{ij} \) that are carried by the nodes at each hop, and the corresponding routing probabilities \( \rho_j = \lambda_{ij} / \lambda_i \).

Tor: In general, a Wardrop equilibrium can be found by solving a convex optimisation problems (see, e.g., [6]), where the Frank–Wolfe method and its extensions are suggested for finding equilibria. There is a separate variable in this optimisation problem for every feasible route (more than \( 10^9 \) in our numerical example). If the nodes can be partitioned in a limited number of “node types” that serve the same hops, we propose a more efficient solution method. We do so by expressing that (i) the latency of all nodes that are used at a hop is fixed, (ii) the arrival rate at each hop needs to match \( \Lambda \), and (iii) no other nodes offer a better latency at each hop. For clarity of exposition, we limit the discussion to the Tor network below, though the method extends to a general (partitioned) network.

For Tor, there are but three hops, and nodes are either gate nodes (g) which can be used at hops 1 and 2, exit nodes (e) which can be used at hops 2 and 3, or normal nodes (n) which can be used at hop 2. There are also combined gate-exit (a) nodes which can be used at all hops.
With a slight abuse of notation, let $\Omega_i$ denote the set of $\alpha$-nodes, with $x \in \{g, e, n, a\}$. Moreover, let $\mu_x$ denote the aggregated service capacity of all $\alpha$-nodes, $\mu = \sum_{x \in \{g, e, n, a\}} \mu_x$.

We first establish the stability condition. First, the combined capacity of all nodes should exceed $3\Lambda$ as all cells enter nodes three times. Secondly, the combined capacity at hops 1 and 3 should exceed $2\Lambda$ as each cell needs an entry and an exit node. Finally, there should be sufficient exit and entry capacity. Summarising, we find the following stability condition,

$$\Lambda < \min \left( \frac{\mu_g + \mu_e + \mu_n + \mu_a}{3}, \frac{\mu_g + \mu_e + \mu_n + \mu_a}{2}, \mu_g + \mu_e + \mu_n + \mu_a \right).$$

The construction of this condition shows that it is necessary. The condition is also sufficient as one can easily construct a stable policy if it is satisfied by reserving the hop 1 and hop 3 capacity first from the $g$-nodes and $e$-nodes, and then adding capacity from the $n$-nodes if required. The remaining capacity can then be used for hop 2.

We now turn to the calculation of the Wardrop equilibrium. Let $\lambda_i(W)$ denote the arrival rate at node $i$ that yields expected latency $W$. Solving Equation (1) for $W$, we find,

$$\lambda_i(W) = \frac{2\mu_i}{W \mu_i - 1}, \quad (2)$$

for $W \geq \mu_i^{-1}$. We further set $\lambda_i(W) = 0$ for $W < \mu_i^{-1}$ when the node cannot deliver the latency. Moreover, for $x \in \{g, e, n, a\}$, let $\Lambda_i(W)$ denote the aggregate arrival rate at all $\alpha$-nodes that yields latency $W$,

$$\Lambda_i(W) = \sum_{x \in \Omega_i} \lambda_i(W).$$

The search for a Wardrop equilibrium can now be reduced to 6 scalar root finding scenarios. These scenarios enumerate the different possible configurations of node-types used at the different hops. If a node is used at different hops, the expected cell latency at these hops must be equal in equilibrium. We then use the calculations above to find the cell latency for the given arrival rate, and check that no better routes are available.

1. All types of nodes are used at hop 2. In this case, there is equal latency at all nodes: $W = W_1 = W_2 = W_3$. As the arrival rate in all hops combined equals $3\Lambda$, we solve $\Lambda_3(W) = \Lambda_2(W) + \Lambda_1(W) = \Lambda_2(W) = a(W)$. This corresponds to a Wardrop equilibrium if and only if $\Lambda_3(W) = \Lambda_2(W) = \Lambda$ and $\Lambda_1(W)$ is $\geq \Lambda$. These inequalities ensure that there is sufficient capacity at each hop. If these equalities do not hold, there is no Wardrop equilibrium with equal latency in all (used) nodes.

2. Only $n$- and $g$-nodes are used at hop 2. In this case, the latency at hop 1 and 2 should be equal, $W_1 = W_2$, and the latency at hop 3 needs to exceed $W_3$ such that one cannot improve in hop 1 (hop 2) by using $a$-nodes $(a$ and $e$-nodes). As the arrival rate in hops 1 and 2 combined equals $2\Lambda$ and the arrival rate at hop 3 equals $\Lambda$, we solve $\Lambda_3(W) + \Lambda_2(W) = 2\Lambda$ for $W$ and $\Lambda_3(W) + \Lambda_2(W) = \Lambda$ for $W_3$. This solution corresponds to a Wardrop equilibrium provided $W < W_3$ and $\Lambda_3(W) \geq \Lambda$. The latter inequality again ensures that there is sufficient service capacity at each hop.

3. Only $n$- and $e$-nodes are used at hop 2. In this case, the latency at hop 2 and 3 should be equal, $W_1 = W_2 = W_3$, and the latency at hop 1 needs to exceed $W$. As the arrival rate in hops 2 and 3 combined equals $2\Lambda$ and the arrival rate at hop 1 equals $\Lambda$, we solve $\Lambda_3(W) + \Lambda_2(W) = 2\Lambda$ for $W_2$ and $\Lambda_3(W) + \Lambda_2(W) = \Lambda$ for $W_3$. This solution corresponds to a Wardrop equilibrium provided $W < W_3$ and $\Lambda_3(W) \geq \Lambda$.

4. Only $n$-nodes are used at hop 2 and $a$-nodes are used at hops 1 and 3. In this case, the latency at hop 1 and 3 should be equal, $W_1 = W_3$, and the latency at hop 2 is smaller than $W$. As the arrival rate in hops 1 and 3 combined equals $2\Lambda$ and the arrival rate at hop 1 equals $\Lambda$, we solve $\Lambda_3(W) + \Lambda_1(W) = 2\Lambda$ for $W_2$ and $\Lambda_3(W) = \Lambda$ for $W_3$. This solution corresponds to a Wardrop equilibrium provided $W_2 < W$, and $\Lambda_3(W) + \Lambda_2(W) \geq \Lambda$ and $\Lambda_3(W) + \Lambda_2(W) \geq \Lambda$.

Evaluation of the scenarios above yields the latencies $W_1$, $W_2$, and $W_3$ at the hops and therefore also the latency through the network $W_1 + W_2 + W_3$. In addition, the arrival rates at all nodes are found by evaluating $2$ in the appropriate latency. The corresponding routing probabilities can then be found by distributing the arrival rate at the nodes over the different hops. Note, however, that these routing probability vectors are not unique.

For example, for scenario 3, $g$ and $a$ nodes are used only at hop 1, so we set $p_1 = \lambda(W)A_1^{-1}$ for $i \in \Omega_2 \cup \Omega_3$ and zero elsewhere. In contrast, $e$-nodes are used at hops 2 and 3, the hop 2 arrival rate at $e$-nodes being $\Lambda = \Lambda_3(W)$ for $i \in \Omega_3$ and zero elsewhere. At hop 2 we can, for example, choose $p_2 = \lambda(W)A_2^{-1}$ for $i \in \Omega_2$, and zero elsewhere. A similar approach applies to the other scenarios.

**Price of anarchy:** Figure 1 plots the optimal latency $W_{opt}$ of [4], the latency in Wardrop equilibrium $W_{Wardrop}$ and the latency with proportional routing $W_{pro}$ versus the load for two snapshots of the Tor network. The load is the fraction of the total network capacity which is used. It is immediately apparent that for the 2020 snapshot, the network capacity cannot be fully used. Close examination of the stability condition reveals that this is caused by a lack of exit capacity.

We can also calculate the price of anarchy (PoA) $W_{Wardrop}/W_{opt}$ which expresses the cost that is paid for the protection against malicious users. Comparing these latencies, we obtain a PoA of approximately 2, for loads above 0.2 in both scenarios, and slightly larger values for lower loads (up to 2.15). Further experimentation showed that these observations were consistent with other snapshots of the Tor network.

**Discussion:** Our results suggest that the Wardrop equilibrium concept can be used to design circuit selection algorithms. Our approach offers two substantial improvements over the optimal assignment of [4]. First, the Wardrop assignment offers protection against latency optimising users as no user can get a better latency by unilaterally changing routes. Second, the equilibrium assignment can be computed very efficiently as at most six root finding problems need to be solved. One can further expect that only one or two root finding problems need to be solved in practice, as the historically most likely scenarios can be evaluated first. Moreover, when recalculating the assignment, it is very likely that the scenario which was applied in the previous assignment, still applies. These advantages do not come for free though. There is a cost for the additional protection in terms of an increased latency. Somewhat
surprisingly, we found that the latency consistently doubles compared to the optimal assignment, which is considerably better than the latency of the proportional assignment of Tor.

Finally, our queueing model only approximates the dynamics of the Tor network and is just a first step towards Wardrop-type circuit selection. The practical implementation of such circuit selection requires additional research, and extensive testing in a network simulation environment. This is part of future work.

© 2021 The Authors. *Electronics Letters* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Received: 26 January 2021  Accepted: 3 March 2021
doi: 10.1049/el.2021.0150

---

**References**

1 Dingledine, R., et al.: Tor: The second-generation onion router. In: USENIX Security Symposium. USENIX Association, Berkeley (2004)
2 Li, B.D., et al.: An overview of anonymity technology usage. *Comput. Commun.* 36, 1269–1283 (2013)
3 Imani, M., et al.: Modified relay selection and circuit selection for faster Tor. *IET Commun.* 13, 2723–2734 (2019)
4 Herbert, S.J., et al.: Optimising node selection probabilities in multi-hop M/D/1 queuing networks to reduce latency of Tor. *Electron. Lett.* 50, 1205–1207 (2014)
5 Takács, L.: *Introduction to the Theory of Queues*. Oxford University Press, Oxford (1962)
6 Correa, J.R., Stier-Moses, N.E.: Wardrop equilibria. In: Encyclopedia of Operations Research and Management Science. Wiley, New York (2010)