

Scalable Resilient Overlay Networks Using Destination-Guided Detouring

Sameer Qazi and Tim Moors

sameerq@student.unsw.edu.au, t.moors@unsw.edu.au.

Abstract— Routing policies used in the Internet tend to be restrictive limiting communication between source-destination pairs to one route, when often better alternates exist. To avoid route flapping, recovery mechanisms may be dampened, making adaptation slow. Unstructured overlays have been widely used to mitigate the issues of path and performance failures in the Internet by routing through alternate paths via overlay peers. The construction of such routing overlays often does not take into account the physical topology of the network which necessitates that all overlay nodes use aggressive online algorithms for end-to-end path discovery, limiting scalability. In this paper, we analyze a topologically-aware architecture to estimate end-to-end path availability for service on Internet. We propose Destination-Guided Detouring via Resilient Overlay Network (DG-RON); a distributed coordinate based overlay architecture, which uses landmark based heuristics for scalable end-to-end path discovery. Simulations show that DG-RON can predict alternate paths with a high success rate.

I. INTRODUCTION

The Internet seems to work most of the time but sometimes recovery from failures is painfully slow. For many of the user perceived performance failures/faults there is a redundant path available which can be used to actually prevent or “mask” the fault from the end user using quick switch-over mechanisms. One study [1] shows that for almost 80% of the paths used in the Internet there is an alternate route with a lower probability of packet loss. The Internet finds alternate paths using the Border Gateway Protocol-4 (BGP-4), where each sub-network learns about global reachability to different hosts in the network through exchange of route advertisements with only immediate neighbors (sub-networks). Despite being highly scalable, there are three fundamental shortcomings with the way alternate paths are explored by BGP. (1) On detection of a failure on the primary route, path exploration proceeds using a ‘trial and error’ method investigating each alternate path in turn; (2) The rate at which new routes are learned is artificially dampened to avoid flooding the network with frequent path update messages causing routing table changes, a phenomenon known as “route flapping”; (3) BGP only addresses reachability and does not adequately address more subtle performance metrics such as latency, loss rates and

throughput on paths. The combined effect of these shortcomings is that BGP can sometimes take as much as 30 minutes to recover from some path failures [2].

The development of overlay routing techniques was a systemic approach to counter these major shortcomings of BGP. An overlay is basically a group of participating peers who agree to route traffic amongst themselves on behalf of other participants (peers) to bypass faults observed in the underlay path. However, the ability of an overlay to find good detours in the Internet is primarily dependent on the knowledge of the underlay topology. This is because an overlay link is logical abstraction of several physical links. Two overlay links may seem disjoint at the application layer yet share a link in the underlying IP layer. The shared IP link renders both useless in the event of failure. For example consider the example in Figure 1(a). If link ‘*l*’ fails disconnecting source *S* from destination *D* then assuming each link has unit weight and using shortest path routing, it renders both overlay peers R_1 and R_2 useless for *S* to reach *D* using a single overlay hop as *S* needs *l* to reach R_2 and R_1 needs it to reach *D*. In this case *S* can only reach *D* through R_3 or through the two hop overlay route $S \rightarrow R_1 \rightarrow R_3 \rightarrow D$. This requires that overlay nodes constantly monitor individual overlay peers and overlay links to successfully detour the traffic via an overlay node in the event of failure on the underlay network.

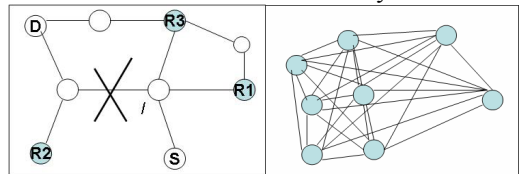


Figure 1(a) (left). How overlay resilience depends on topology of the underlay network. **Figure 1(b)** Inferring maximum information about all virtual overlay links.

Maintaining complete state about all overlay peers and the inter-connecting overlay links require in the ideal case that all ‘*N*’ peers be connected as logical mesh or clique (Figure 1(b)). Subsequent probing for measurement of end-to-end path metrics and its dissemination via link state protocol incurs maintenance overheads of $O(N^2)$. This results in scalability issues limiting the size of deployed overlay networks. On the contrary, maintaining complete overlay state without the knowledge of topological diversity of individual relay nodes may be counterintuitive when we consider that the location of path and performance failures are not known a

priori, are often correlated and vary on very small time scales. Current works, e.g. RON [3] aim to bypass path failures using application specific metrics e.g. throughput, loss rate, latency and routing through all possible indirect overlay nodes which are probed aggressively incurring large overheads. Such path exploration techniques are not scalable above modest network sizes.

Contribution: To address failures in the Internet adequately, we have three major requirements; scalability, quick availability of alternate paths, and ability to find paths based on application specific metrics. BGP only addresses scalability; RON[3] addresses BGP shortcomings adequately but lacks scalability. We propose a new scalable and topology-aware architecture DG-RON (Table I) in which overlay nodes organize themselves into a logical coordinate space with the help of landmarks in a hierarchy simplifying the path exploration problem for packet detouring requests through use of offline path searching mechanisms. Our work is different from RON [3] in that we choose alternate paths only when the default paths are deemed as failed, instead of switching paths based on minor performance gains as such schemes are not scalable. Simulation results show that good predictions can be made for estimation of overlay path availability for desired performance using low probing costs even when the overlay peers maintain overlay state asynchronously.

Table I. Optimizing search for alternate paths in large networks.

	BGP	RON[3]	DG-RON
Scalability	√	X	√
Failure Detection Speed	X	√	√
Detecting Performance Failures	X	√	√

The main contributions of this paper are: (1) To see how a prediction-based approach can lead to scalable end-to-end path discovery while lowering path monitoring overheads; (2) Use a bottom-up approach in designing a topology-aware architecture for resilient overlay network with scalability on Internet proportions; (3) Present landmark based heuristics for overlay peers to select alternate paths. The paper is organized as follows: Section 2 discusses related work; Section 3 presents design heuristics. In Section 4 we evaluate the performance of the proposed Destination Guided Detouring Service and finally Section 5 presents the conclusion.

II. RELATED WORK

Anderson et al, designed RON [3] to be a resilient routing tool for the Internet by implementing a small link state overlay (50 nodes). The overlay tries to find the best alternate path to the destination. The best path may be the default Internet path or an alternate overlay path. The design posed scalability problems with more than 50 nodes due to extreme bandwidth requirements for active probing of all virtual overlay links and subsequent dissemination of the learned parameters over the logical mesh architecture, at intervals of the order of a few

seconds. [4] shows that in most cases alternate paths can be found using at most one overlay hop but does not address the path selection problem. Several other studies including [5] have addressed the scalability issue in unstructured overlays by arguing in favor of a hybrid architecture, combining overlay routing with multi-homing techniques. However, Multi-homing is only effective in improving path diversity near the edge of the network and thus has limited benefits for failures other than last hop failures. Multi-path routing on overlay networks is also proposed in [6].

Topology aware approaches have been extensively studied to counter the scalability issue. [7] discusses a proposal for ‘pruning’ the overlay topology through removal of redundant physical links from the path monitoring exercise which are not likely to be selected by the overlay routing algorithm.

Several other studies present interesting heuristics for finding disjoint overlay paths without explicit knowledge of underlay topology or aggressive path monitoring. Akamai driven one hop source routing is presented in [8] which presents a detour selection process in which a host selects as detour, an overlay node in physical proximity to one of the ‘preferred’ Akamai servers (mirrors) to serve content. [9] presents a scheme where a source uses routing tags, each of which specify the path a packet takes through the network from its present location to the destination by selecting one of possible routing options.

III. ARCHITECTURE

A. DG-RON Clients and Services

We assume that DG-RON clients subscribe to the service from the nearest DG-RON edge node stipulating services required e.g. connectivity to popular destinations but use the services on ‘pay per use basis’ where packet detouring requests are only made once the default path suffers a performance or path failure. This is to ensure that overlay based path switching does not affect non-overlay traffic or cause oscillations by frequently swapping paths for minor performance gains [10]. We assume that the packet to be routed enters the overlay via its nearest overlay proxy after encapsulation and departs at another depending on the path selection algorithm.

B. Overlay Infrastructure

The purpose of a resilient routing overlay is to provide improved connectivity between any two arbitrary hosts on the underlay network in the face of failures. Such a service should be scalable, provide satisfactory performance guarantees, be able to handle overlay churn and provide good load balancing on physical links. Keeping these global objectives in mind we start with a bottom-up approach in overlay construction.

BGP has taught us the importance of hierarchy for global scalability. We choose to use architectural hierarchy to meet this objective. The architecture uses ‘n’ landmarks to divide the overlay network into ‘n’ logical zones and at the same time into an ‘n’ dimensional co-ordinate space for inter-host

distance estimation (Figure 2). In our simulations we set $n=7$ which results in optimum results for inter-host distance estimation [11, 12]. The landmarks could become a potential performance bottleneck in the system so a single landmark could actually be a logical abstraction of a group of machines collocated together or in close proximity of each other [13]. Each of the landmarks is responsible for the bootstrapping of new peers in its own logical zone. Landmark nodes only play a role in forming the infrastructure of the overlay but do not participate in routing.

Each overlay node measures its distance from each of the landmarks as round trip time (in milliseconds) between a ping request and reply; and stores the result as an n -dimensional network vector $[rtt_1 \ rtt_2 \ rtt_3 \ \dots \ rtt_n]$ (where n is the number of reference landmarks used in simulation). Each overlay node then contacts its nearest landmark node to join its logical zone and to request a detour set. The members in the detour set of each overlay node are selected in the DG-RON architecture using the ‘binning’ technique proposed in literature [13]. Each peer requests a total of T relay nodes from its nearest landmark (as explained previously). In this peer selection method a landmark returns x short distance (intra-zone) peers from its own logical zone, and the remaining $y (=T-x)$ are long distance (inter-zone) peers requested from other landmarks. The distance estimation function used by landmarks is similar to the Cartesian Distance estimation method ‘IP2GEO’ [11]. The network distance is estimated between the network vectors of different nodes in different zones for each of the landmarks and the network vector of the requesting node. The network distance in terms of rtt metric between two arbitrary nodes ‘a’ and ‘b’ is estimated from their network vectors as shown by equation below.

$$Dist = \sqrt{\{(rtt_{a1} - rtt_{b1})^2 + (rtt_{a2} - rtt_{b2})^2 + \dots + (rtt_{an} - rtt_{bn})^2\}}$$

(where rtt_{an} is the round trip time of node ‘a’ in milliseconds from landmark ‘n’)

Selecting relay nodes using the binning heuristic ensures that the overlay connectivity is maintained and the average routing latency on the overlay is low [13].

Once a peer obtains its detour set from the landmark it probes this detour set. We propose a flexible randomized probing scheme. At each probing epoch, every overlay participant probes one random peer in its detour set. This is different from [3] where each overlay peer probes each of the other overlay peers using probe timers and a round-robin like scheme and detection of failed peers using time-outs. As we highlighted earlier, the motivation of our design is scalability at Internet proportions. The randomized probing scheme does not require that overlay peers probe aggressively or to implement timers and timeout functions. If a peer is deemed as failed for a considerable time interval then a new peer can be eventually requested from the landmark. However, in the simulations we do not implement any repairs to the detour set of the overlay nodes and investigate only the static resilience of the overlay using only the live members of the detour set. Such assumption is reasonable in real deployment of DG-RON with non-aggressive probing epochs. The landmark

based decentralized architecture eliminates the need for any information flooding in the network as required in link-state protocols making the design scalable for large overlay networks.

C. Selection of Alternate Paths

We propose scalable offline mechanisms to find alternate paths where we do not need to address the actual composition of the original underlay path suffering from a performance failure event. Online link probing techniques such as those used by [3] are still required for performance measurements to determine dynamic performance; however such overheads are significantly reduced owing to the distributed architecture. The offline heuristics for selecting topologically diverse detours only require that the destination be mapped into the network co-ordinate space. This mapping can easily be managed by the landmarks for popular destinations to which DG-RON clients have subscribed. For unfamiliar destinations the landmarks could extrapolate the approximate co-ordinate vector using vectors of other peers from its nearest landmark optionally utilizing services of a third party e.g. WHOIS servers. Only the knowledge of the destination IP address is required for both and should suffice to find the overlay based detour. Such information could also be cached as frequent requests to popular destinations are made so the peer can incrementally learn about these. Next we describe three offline schemes for selection of an overlay ‘detour’ node once the position of the destination has been determined in the co-ordinate space.

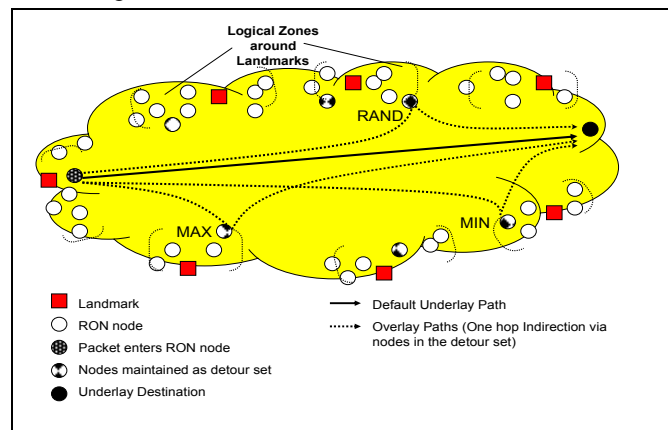


Figure 2. Finding Topologically diverse detours for underlay destinations.

1) Selecting Detours

[3, 4, 8] show that in most cases a performance failure can be bypassed using single hop indirection using an overlay node. We use the ‘Maximum Divergence Heuristic’ to find such one hop detours, in which the peer chooses next hop based on host distance (Cartesian distance) of the destination from the eligible next hop candidate relay nodes. The underlying idea is not unlike the ‘Earliest Divergence Heuristic’ [14], which aims to select a path which diverges from the default path near the source and converges near the destination in order to avoid a failed or congested link. However, [14] assumes the availability of complete AS path information between the source-destination pair and candidate

alternate paths. We argue that such information is seldom known in advance and maintaining it would again lead to scalability issues that we aim to address in this paper. Our ‘divergence’ criteria is to select with good probability an alternate path that diverges from the defunct portion of the default path near the location of the failure, e.g. a congested link. Eligibility of such overlay paths may further be based on underlying network characteristics, e.g. loss rates, latency or throughput through monitoring (as explained in the next section). The first heuristic we use in searching for such divergent paths is MAX, in which we choose an overlay peer which has the maximum network distance from the destination. The underlying reason for using MAX is to select with high probability an overlay peer which leads to a topologically diverse alternate path to reach the destination. We also search for alternate paths using MIN where we use overlay peers close to the destination as detours. The underlying heuristic for this rule in contrast to the previously mentioned MAX rule is the observation of fact that many paths in the Internet violate the triangle inequality[15]. Thus, it is also possible to find a disjoint path using a peer in proximity to the destination. Also, last hop failures are avoidable using a detour close to the destination with a high probability. We also observed, that since Internet has a power-law topology [16] landmark based decisions may not always work in finding disjoint paths, so we also use a random selection rule, RAND, in which a peer arbitrarily selects an eligible live peer as next hop. Figure 2, shows the underlying idea in the selection of detours. There may be other landmark based heuristics which we may have neglected in our paper and may work better than the ones presented in the paper; our main objective is to investigate if such schemes can work in selection of disjoint paths when the cause and location of the path failure on the primary path is not known in advance. The generic algorithm for offline detour selection is presented in Figure 3.

2) Overlay Link Monitoring

The offline methods based on network co-ordinates (discussed in the previous section) can embed only latency but not failure or congestion information; and thus may not adapt well for dynamic performance estimation on alternate paths.

INPUT: Network Coordinates for destination D and detour set $T = T_1, T_2, \dots, T_n$.

OUTPUT: Candidate Alternate paths for destination D

Algorithm: (FIND_CANDIDATE_ALTERNATE_PATHS)

*Define: Cost = Distance($v(A), v(B)$)
(where: $v(A)$ and $v(B)$ are network vectors for any arbitrary hosts A & B, in the network co-ordinate space)*

*Repeat for $i=1$ to $|T|$
Cost (i) = Distance ($v(D), v(T_i)$)*

*Cost_A = min[Cost (i)] : $i, A \in [1, 2, \dots, |T|]$
Cost_B = max[Cost (i)] : $i, B \in [1, 2, \dots, |T|]$*

*If MIN, NextHop = T_A
If MAX, NextHop = T_B
If RAND, NextHop = T_{rand}*

Figure 3. Selection of next hop based on ‘Maximum Divergence Principle’.

Thus, to supplement the offline path selection process online path monitoring is necessary in DG-RON. Overlay links between an overlay peer and peers in its detour set are monitored for performance characteristics such as latency, throughput and loss rates. Note that we only probe overlay links to candidate detours; [8] shows that predicting good detouring points can yield acceptable upper-bounds for end-to-end path metrics. We conjecture that the underlying reason for this is the small probability for many internet links on spatially diverse paths to undergo congestion at similar points in time. Moreover, unlike [8] we combine the disjointness criteria with absolute performance merits to optimize the selection of candidate detouring nodes. To improve scalability further, we propose that probing could be replaced by passive monitoring of traffic traversing overlay links between a peer and its detour set to improve dynamic estimation of path performance without introducing any probing traffic and subsequent probing overheads. Techniques for both active network probing and passive traffic monitoring have been studied in the past e.g. [3, 17] and is beyond the scope of our present work.

IV. PERFORMANCE EVALUATION

Several works [3, 6, 18] have investigated resilience of wide area routing overlays for masking performance or path failures in the underlay network using Planet Lab¹ test bed using active monitoring or analyzing offline data from such studies. While such studies are attractive in that they help capture actual performance data from the Internet, the data from such studies can be biased when we want to investigate the true extent of topological path diversity. For example most RON hosts in [3] are based in North America with many participating from educational institutions and [18] deliberately did not select hosts outside North-America as “...many alternate paths to these sites would, in most cases, share the same transoceanic link and would have provided less chance for performance gain”. Moreover, landmark-based mapping is an essential component of our architecture, due to which we use simulations to investigate the design heuristics proposed earlier.

A. Simulation Methodology

Internet has been popularly represented as AS graphs or router level graphs $G = (V, E)$ where the vertex set V represents the different autonomous systems or hosts (routers) and the edge set E representing the peering (routing) relationships between these ASes or hosts. There are several popular power law Internet topology generators [16], e.g. Inet², BRITE³; however, these models capture only the large scale inter-AS level structure but not the intra-domain connectivity which is essential for the investigation of path diversity. Transit-Stub topologies generated using GT-ITM⁴

¹ PlanetLab.<http://www.planet-lab.org/>

² University of Michigan. Inet 3.0. <http://topology.eecs.umich.edu/inet/>

³ Boston University BRITE. <http://www.cs.bu.edu/faculty/matta/Research/BRITE/>

⁴ Georgia Tech GT-ITM topology generator. www-static.cc.gatech.edu/projects/gtitm/

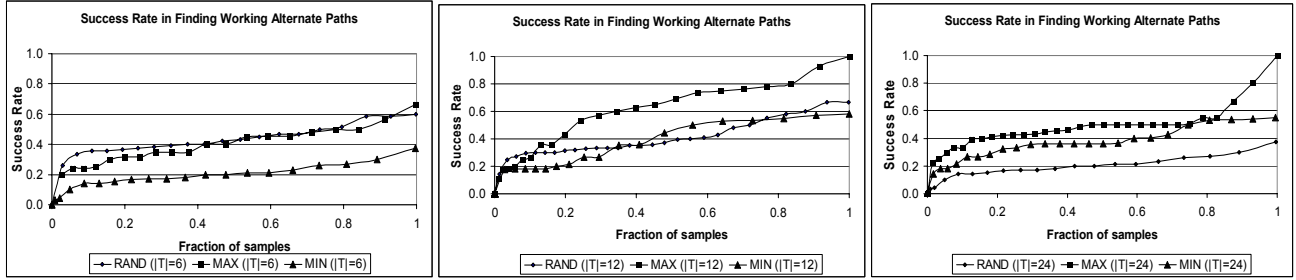


Fig 4. Success Rate in finding alternate paths using RAND, MAX and MIN (a) $|T|=6$, (b) $|T|=12$ & (c) $|T|=24$.

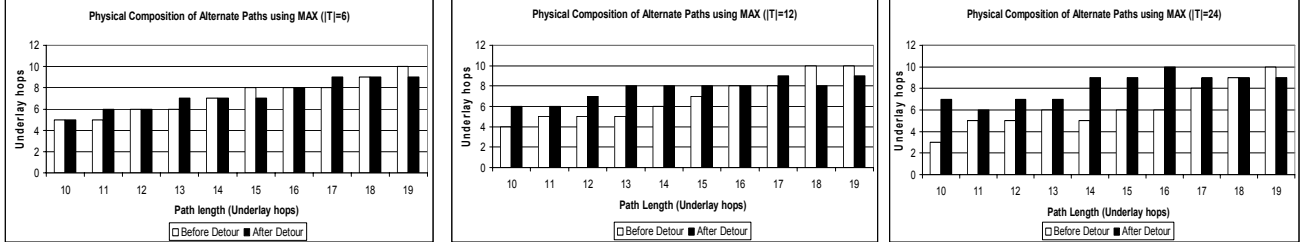


Fig 5. Physical Composition of Alternate paths found using MAX for (a) $|T|=6$, (b) $|T|=12$ & (c) $|T|=24$.

on the other hand can model both the intra-domain and inter-domain connectivity.

1) Generation of Underlay Topology

We use Internet-like transit-stub topologies in our simulations for the generation of the underlay network. We use several 600 node transit-stub topologies generated using GT-ITM. Note the node count includes both routers and end hosts. Each graph is made up of three transit domains with each transit domain consisting on average of eight transit nodes. The transit nodes contain edges amongst themselves with a probability of 0.5. Each transit node has three stub domains. Each stub domain consists of an average of 8 nodes, in which nodes are once again connected with a probability of 0.5. These parameters are from the sample graphs in the GT-ITM distribution; we are unaware of any published work that describes parameters that might better model common Internet topologies.

2) Selection of Landmarks and Overlay Nodes

Overlay nodes are usually located at edge of the network and selection of such edge nodes in simulated topologies is usually based on the node out-degree where nodes with high out-degree (large number of links) represent the ‘rich club’[16] in the network core while the edge comprises of nodes with very low out-degree. To meet this objective, we randomly select 7 and 50 edge nodes from the stub domains to be landmarks and overlay nodes respectively. Likewise, for packet detouring requests to DG-RON nodes, underlay destinations are also randomly selected from the edge nodes in the topology comprising of nodes in the stub domains. In all simulations, we keep the overlay size fixed as 50 nodes, since (1) this is typical size of a commercially-deployed routing overlay [3] from literature; (2) all heuristics we present in the paper are applicable for power-law networks like the Internet; and thus are equally valid for overlay networks of other arbitrary size; however as we show later, the performance of

individual heuristics presented may be affected under different situations. For the implementation of the proposed landmark based overlay and routing algorithms we use ns-2⁵.

3) Failure Model

Modeling performance failures in the Internet accurately is an impossible task and is not the primary scope of our work. The purpose of our work is to identify the potential benefits of prediction-based path selection as a scalable packet detouring service. Failures in the underlay network are usually link or router events. While both router and link events may last several minutes owing to BGP convergence delay described previously; performance failures are usually short duration link events and RONS are normally deployed to provide alternate routing. Due to the limited access to the information of link failure patterns in the real networks, we assume the conditions on links vary between being normal or lossy [19] and use an exponential link-failure model, assuming that both up-times and down-times of a physical link follow exponential distributions; and set the failure probability as 10^{-5} . Also, some ‘bad’ links may be more prone to such failures than others [17]; we pick a random 2%-5% links and assign high failure probabilities than others (10^{-3}). Probing Epochs in the overlay are set to a constant rate which is half the mean duration of failures on high failure links in the exponential failure model.

B. Impact of the size of Detour Set

To evaluate the performance merits of each landmark-based path searching heuristic, we conduct a series of simulations to find the optimum size of candidate detour set to get the maximum benefit. To see the impact of the size of detour set each peer maintains based on the proposed detour selection criteria, we find that it is essential for the detour set to capture maximum topological diversity on alternate paths possible. We compare the efficiency of each heuristic in finding alternate paths when RON size $N=50$ and $|T|=3, 6$ & 12 (Fig 4(a-c)) by plotting the CDF of success rate with which each

⁵ The Network Simulator ns-2. <http://www.isi.edu/nsnam/ns/>

heuristic can find alternate paths. We define success rate as the total fraction of alternate paths found using each individual packet detouring heuristic for 1000 source-destination pairs when the underlay path experiences a failure in each simulation.

We see that when the detour set is small $|T|=6$, the performance of RAND is best with a success rate 0.4 in a majority of our samples. MAX similarly has success rate 0.4 in 60% of our samples; MIN has a success rate of only 0.20 in 70% of the total samples. This is because the detour set is too small to exploit the benefits of landmark based detouring.

Increasing $|T|$ to 12 and 24 visibly degrades the performance of RAND showing random selection of a spatially diverse detouring point may not be good when the detour set is sufficiently large. Nevertheless, our initial suspicion that landmark based heuristics may not always work for power-law topologies is confirmed by the observation that RAND is able to find a sufficient number of alternate paths in all cases.

Increasing $|T|$ to 12 improves the performance of MAX which now has a success rate of > 0.6 in discovery of alternate paths in 60% of samples but further increase to $|T|=24$ slightly reduces its performance. The cause for the degradation with increasing detour set size is the observation that as MAX starts gradually picking as detour, an overlay node, which is farther from the destination in the landmark space, it increases the chances of selecting an overlay node closer to the source (Figure 5). For instance, the detouring overlay node in a 16 hop path is roughly 8 underlay hops from both the source and the destination when $|T|$ is 6 (Figure 5(a)) but only 6 hops from the source when $|T|$ is increased to 24 (Figure 5(c)). This increases the probability that the overlay path will overlap with the default underlay path and thus not able to avoid the failed link.

We see similar pattern in the case of MIN (graphs not shown) which has slightly worse performance when $|T|=24$; success rate of 0.4 or under for 70% of our samples compared with only 40% samples when $|T|=12$. Picking an overlay node closer to the destination increases the probability that the path may overlap with the default end-to-end path and thus go through the failed link.

In summary, the experiments reveal that: (1) In most cases alternate paths have a detouring node which is roughly midway along the alternate path showing that Maximum Divergence is possible through landmark-based techniques;

(2) The detour set does not need to be very large but must capture topological diversity in the network sufficiently; (3) Each offline technique is able to address individual failures in the space domain uniquely and thus, complement each other when the failure on the underlay path is not known a priori. Using a combination of such techniques for path exploration can harness the RON sufficiently to deal with individual failures.

Table 2. Physical Characteristics of Paths found by MAX, MIN and RAND ($|T|=12$).

Heuristic	Path Length (Underlay Hops)	Underlay Hops to Detouring Overlay Node	Underlay Hops from Detour to Destination	Path Inflation (Hops)
MAX	13.57	6.23	7.33	1.85
MIN	10.33	6.17	4.17	1.34
RAND	13.42	6.16	7.26	1.75

C. Cost of Packet Detouring via DG-RON

We analyze the cost of packet detouring via DG-RON through measurement of path stretch encountered on alternate paths. We first consider the physical characteristics of alternate paths found by each heuristic. Table 2 summarizes our results. We note that alternate paths found by MAX and RAND on average traverse 13+ network hops and those found using MIN are 3 hops shorter. All successful detours found using MIN are on average 3 hops closer to the destination than those found using MAX and RAND. The relative delay penalty (RDP) and relative hop penalty (RHP) incurred on alternate paths in terms of physical network hops and latency is 1.65 and ~ 2 respectively.

D. Comparison with RON-like schemes

We compare the performance-tradeoffs of DG-RON with a RON-like system by observing the fraction of underlay failures successfully masked by both schemes. RON-like uses a full detour set ($|T|=50$) and searches for all live one-hop paths post-detection of a failure. Note that we don't attempt to find multiple alternate paths in both DG-RON and RON-like schemes and terminate the search once a single alternate path has been found successfully (not extending the search beyond the 3 candidate detouring nodes in DG-RON). This is because many studies e.g. [20] point out that it suffices to find just a single alternate path to match the QoS requirement as Internet flows seeking QoS guarantees only attempt to optimize a single parameter out of throughput, loss rate or latency; this can be met by discovery of just a single path with high probability. This is also to meet our initial objective of a

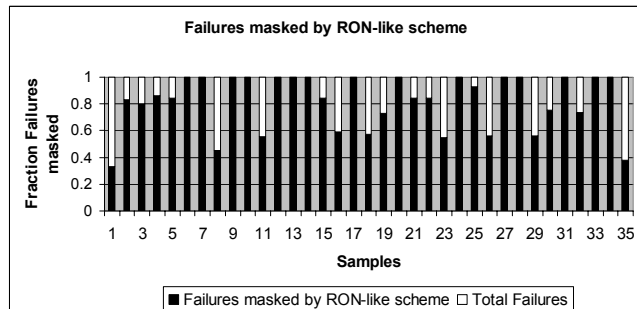
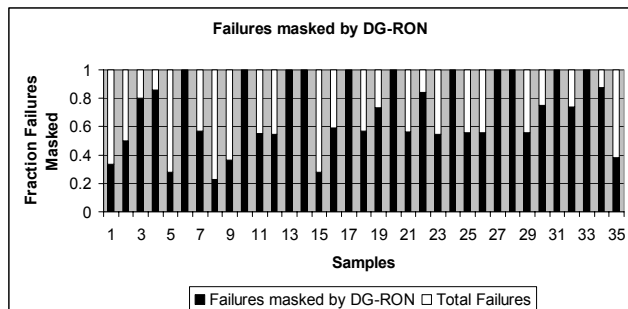


Figure 6. Success Rate of DG-RON and a RON-like scheme in masking failures.

globally-scalable best-effort RON service. Also, note while DG-RON uses randomized probing as explained previously RON-like uses probing like MIT-RON [3] where each RON participant probes virtual overlay links to each of the other RON nodes in round-robin style at beginning of every probe interval. We neglect all cases where there existed no alternate one-hop paths via RON for a more meaningful comparison. Figure 6 shows the results obtained. Even though, we see a wide variability in the results, it is clear that DG-RON masks a significant fraction of the failures which is possible through a RON-like scheme. When a small percentage of the physical links suffer from performance failures in the underlay network, RON-like schemes only outperform DG-RON in 30% of the samples where percentage of failures masked was 60% or higher. DG-RON is able to mask all failures possible using RON-like schemes in $\sim 70\%$ samples. Note large performance gaps between the two schemes occur mainly as a result of small number of failures seen by both schemes in the simulation. Moreover, performance benefits of the RON-like approach closely resemble those reported by RON [3]; which show that mesh-based overlays are effective in masking 60-100% of failures in most cases. This shows that the resilience of the overlay is constrained by spatial distribution of failures giving credence to the claim in [21] that overlay network without effective multi-homing may not be able to provide high degree of reliability and performance gains in all cases.

V. DISCUSSION

The simulation results presented in the previous section reveal that landmark based offline path searching methods can work well in power-law topologies such as the Internet which can supplement or reduce the overheads of aggressive online, path selection algorithms. Although, the results presented in this section do not show the absolute performance merits of DG-RON; it does imply that using a topology-aware approach there is ample opportunity for finding alternate paths even if overlay peers are not connected as full mesh. Considering that different routing domains are administered and configured independently; and performance failures are short duration events, making it highly unlikely for a large fraction of links to undergo congestion or suffer from other performance degradations at the same time, DG-RON can predict good alternate paths among candidate nodes in the detour set.

The proposed design does have some obvious caveats; the most glaring of all is the fact that the path exploration could incur some delay in alternate path discovery. We argue that to achieve scalability, this problem is unavoidable. BGP has taught us that scalability only results by marching through all possible alternate paths post-detection of a failure. The landmark based architecture can effectively predict availability of good alternate paths.

VI. CONCLUSION

As the Internet continues to grow, so does the diversity of the connectivity between the nodes. In this paper we seek

insight into scalable overlay-based techniques for discovering infrastructural redundancy and robustness potentially present in the Internet. Current solutions e.g. RON[3] unnecessarily search through a large path exploration space and subsequent overheads associated with aggressive path monitoring pose scalability issues. To address this issue several previous works [14, 22] have focused on topology aware heuristics in overlay construction and link monitoring which make it possible to both monitor and select alternate paths using distributed approaches. Our work is similar to such approaches in that we aim to lower both path monitoring overheads and reduce the candidate path exploration space. In addition our work presents a platform for harnessing the findings of previous literature [4, 14].

In this paper, we describe an architecture for a best-effort RON service, DG-RON; which simplifies the path exploration problem by finding topologically diverse detours, using small candidate detour sets. We also present three offline heuristics which complement each other under different spatial distribution of failures in finding available paths via DG-RON with a high probability. We show that landmark based heuristics can work well for power-law networks like the Internet for finding topologically diverse alternate paths. Our comparison of DG-RON with RON-like schemes show that it is possible to find alternate paths with a high probability while incurring low measurement and maintenance overheads.

REFERENCES:

1. Savage, S., et al., *Detour: a Case for Informed Internet Routing and Transport*. IEEE Micro, 1999. Vol 19, no 1 p. 50-59.
2. Labovitz, C., et al., *Delayed Internet routing convergence*. Networking, IEEE/ACM Transactions on, 2001. 9(3): p. 293-306.
3. Andersen, D., et al. *Resilient overlay networks*. in *SOSP '01: Proceedings of the eighteenth ACM symposium on Operating systems principles*. 2001: ACM.
4. Gummadi, K., et al. *Improving the Reliability of Internet Paths with One-hop Source Routing*. in *OSDI '04*. 2004.
5. Andersen, D.G., et al., *Improving Web Availability for Clients with MONET*. NSDI'05, Boston MA.
6. Andersen, D., A. Snoeren, and H. Balakrishnan. *Best-path vs. multi-path overlay routing*. in *IMC '03: Proceedings of the 3rd ACM SIGCOMM conference on Internet measurement*. 2003: ACM.
7. Nakao, A., L. Peterson, and A. Bavier, *Scalable routing overlay networks*. SIGOPS Oper. Syst. Rev., 2006. 40(1): p. 49-61.
8. Su, A.-J., et al. *Drafting Behind Akamai (Travelocity-Based Detouring)*. in *SIGCOM 06*. 2006. Pisa, Italy.
9. Yang, X. and D. Wetherall. *Source Selectable Path Diversity via Routing Deflections*. in *Sigcom 06*. 2006. Pisa, Italy.
10. Keralapura, R., et al., *Can ISPs Take the Heat from Overlay Networks?*, in *HotNets (04)*. 2004: San Diego, CA USA
11. Padmanabhan, V.N. and L. Subramanian. *An Investigation of Geographic Mapping Techniques for Internet Hosts*. in *SIGCOMM '01*. 2001. San Diego, California, USA.
12. Ng, T.S.E. and H. Zhang. *Predicting Internet network distance with coordinates-based approaches*. in *INFOCOM 2002. Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*. 2002.
13. Ratnasamy, S., et al. *Topologically-Aware Overlay Construction and Server Selection*. in *Infocom. 2002*. New York, NY, USA.
14. Fei, T., et al. *How to Select a Good Alternate Path in Large Peer-to-Peer Systems?* in *Infocomm 06*. 2006. Barcelona, Spain.
15. Tang, L. and M. Crovella. *Virtual Landmarks for the Internet*. in *IMC'03*. 2003. Miami Beach, Florida, USA.
16. Zhou, S. and R.J. Mondragon, *The rich club phenomenon in internet topology*. IEEE Communication letters, 2004. 8(3): p. 180-182.
17. Padmanabhan, V.N., L. Qiu, and H.J. Wang. *Server-based inference of Internet link lossiness*. in *INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications Societies. IEEE*. 2003.
18. Rewaskar, S. and J. Kaur. *Testing the Scalability of Overlay Routing Infrastructures*. in *Passive And Active Measurements (PAM 04)*. 2004. Antibes Juan-les-Pins, France.
19. Cui, W., I. Stoica, and R.H. Katz. *Backup path allocation based on a correlated link failure probability model in overlay networks*. in *Proceedings of 10th IEEE International Conference on Network Protocols (ICNP'02)*. 2002. Paris, France
20. Subramanian, L., et al., *OverQoS: offering Internet QoS using overlays*. SIGCOMM Comput. Commun. Rev., 2003. 33(1): p. 11-16.
21. Akella, A., et al. *A comparison of overlay routing and multihoming route control*. in *SIGCOMM '04: Proceedings of the 2004 conference on Applications, technologies, architectures, and protocols for computer communications*. 2004: ACM.
22. Tang, C. and P.K. McKinley. *A distributed approach to topology-aware overlay path monitoring*. in *Distributed Computing Systems, 2004. Proceedings. 24th International Conference on*. 2004.

