The ISP Column

A monthly column on things Internet

March 2012
Geoff Huston

Leaking Routes

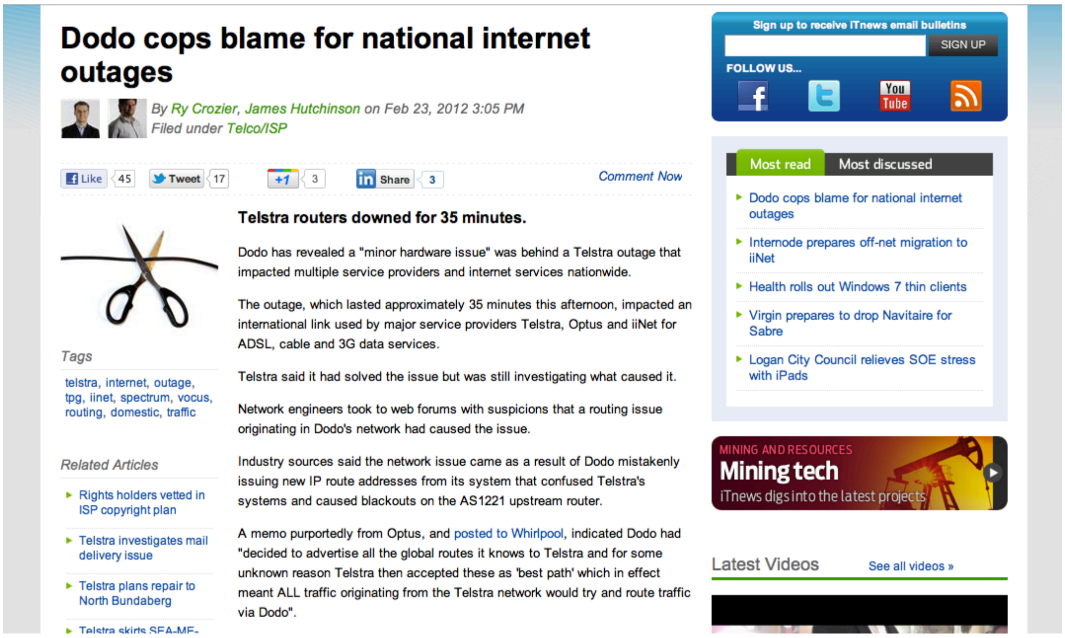
Its happened again.

We've just had yet another major routing leak, this time bringing down the Internet for most of an entire country. Maybe twenty years ago no one would've noticed, let alone comment, but now of course its headline material in the media. What happened? And how could this have been prevented? Can we do better? I'd like to look at this incident in here, and also look at the implications for the current efforts to secure our inter-domain routing system, BGP.

In my previous column I described an approach to detect the presence of so-called bogon filters in the Internet by using online ads, and embedding our reachability tests into the advertisement using embedded Flash code. (http://www.potaroo.net/ispcol/2012-02/bogonfilter.html) In that article I also noted that the continued use of bogon filters were perhaps an anachronism in today world:

> "The continued efficacy in such filters in eliminating malware and abuse appeared to have little in the way of factual substantiation. But the ISP security industry apparently loves a good pantomime, where the superficial veneer of security replaces any substantive and potentially more intrusive and expense security response, and the use of bogon filters was certainly a widespread item of security pantomime in the ISP 's operational manual."

But you shouldn't infer from that comment that all forms of route filtering should be dropped. Indeed the consequences of dropping route filters from external interfaces can be incredibly disruptive, as the following clipping from an Australian IT wire service (http://www.itnews.com.au) illustrates:



Dodo cops blame for national internet outages

By Ry Crozier, James Hutchinson on Feb 23, 2012 3:05 PM
Filed under Telco/ISP

Telstra routers downed for 35 minutes.

Dodo has revealed a "minor hardware issue" was behind a Telstra outage that impacted multiple service providers and internet services nationwide.

The outage, which lasted approximately 35 minutes this afternoon, impacted an international link used by major service providers Telstra, Optus and iiNet for ADSL, cable and 3G data services.

Telstra said it had solved the issue but was still investigating what caused it.

Network engineers took to web forums with suspicions that a routing issue originating in Dodo's network had caused the issue.

Industry sources said the network issue came as a result of Dodo mistakenly issuing new IP route addresses from its system that confused Telstra's systems and caused blackouts on the AS1221 upstream router.

A memo purportedly from Optus, and posted to Whirlpool, indicated Dodo had "decided to advertise all the global routes it knows to Telstra and for some unknown reason Telstra then accepted these as 'best path' which in effect meant ALL traffic originating from the Telstra network would try and route traffic via Dodo".

Like 45
Tweet 17
+1 3
Share 3
Comment Now
Tags
telstra, internet, outage, tpg, iinet, spectrum, vocus, routing, domestic, traffic
Related Articles
Rights holders vetted in ISP copyright plan
Telstra investigates mail delivery issue
Telstra plans repair to North Bundaberg
Telstra skirts SEA-ME-
Sign up to receive iTnews email bulletins
SIGN UP
FOLLOW US...
You Tube
Most read
Most discussed
Dodo cops blame for national internet outages
Internode prepares off-net migration to iiNet
Health rolls out Windows 7 thin clients
Virgin prepares to drop Navitaire for Sabre
Logan City Council relieves SOE stress with iPads
MINING AND RESOURCES
Mining tech
iTnews digs into the latest projects
Latest Videos
See all videos »

http://www.itnews.com.au/News/291364,dodo-cops-blame-for-national-internet-outages.aspx

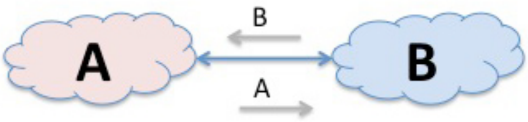
One could be mistaken for being a little confused about route filtering at this point, as it does seem to be a case of dammed if you filter and dammed if you don't! Lets look at this outage in a little more detail and take a guess as to its causes and the possible responses.

The explanation requires some understanding of the way in which the Internet is glued together, so I thought a quick tutorial for those readers who are not completely familiar with the models of peering and interconnection in the Internet might be helpful.

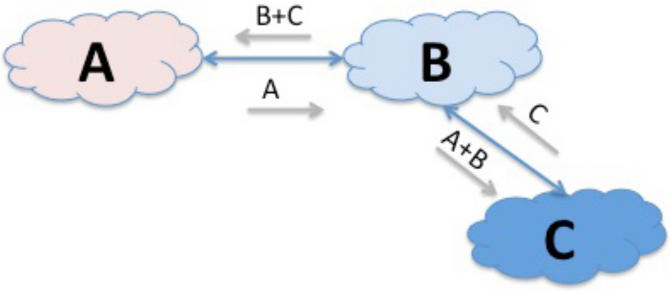
Interconnection and Routing in the Internet – a quick tutorial

When any two networks interconnect, then the way in which they learn about each other is via an exchange of routing information. A "network" in this case can be though of as a collection of reachable IP addresses ("routes"), and the connection of two networks implements a simple forms of "I'll tell you my routes if you tell me yours."

Let's use two networks, and call them A and B. And lets assume that the networks A and B have interconnected in this manner. So now if a source in A's network wants to send a packet to a destination in B's network, then as A and B are directly connected then A has learned all about the set of IP addresses that are reachable in B's network and A's routing system will direct the packet to your network.



Now lets add a third element to this model. What if we have a third network C, that connected to B? If B was prepared to act as a "transit" service provider then if could announce C's routes to A, and announce A's route's to C. Now all points in these three networks can reach each other. If a source in A sends a packet to a destination in C, then A's routing system will direct the packet to network B. B's routing system will recognise the destination address as one located in network C, and will pass it across to C.



Conventionally, we call A, B and C *Autonomous Systems*, and the routing protocol used to exchange routes *BGP*. And if you repeat and rinse the above interconnection scenario another 40,000 times to accommodate the interconnection of 40,000 networks, and use BGP to exchange route information for some 400,000 routes then you end up with something that is much the same scale as today's Internet.

There are many ways one could interconnect 40,000 AS's using the basic tool of pairwise connection. We could use linear connectivity, rings, hub and spoke, or any form of connection topology. What shapes the Internet's particular configuration?

The best answer I can offer is that the Internet is shaped a combination of money and geography.

Geography is the tendency for networks to interconnect to other networks that are physically close. There is a significant industry devoted to running so-called "exchange points" all over the world, which is a dedicated facility where local networks can drop a connection and use the exchange point's switching equipment to interconnect with all the other networks who also present themselves at this point. The motivations for geographic proximity rest in performance and, of course, money. If two networks have an interconnection path that spans the world then the time taken for a packet to traverse this extended path will be far slower than it takes for a packet to traverse a path that spans a metro area or a continental domain. So closer connectivity creates a superior user experience. And, while its not universally true, its certainly more common than not, that longer paths, particularly those that span one or more oceans, cost more per packet than shorter paths. So, in general, shorter network paths are cheaper.

Money is the major motivator for interconnection, as, ultimately its money that drives this industry. To illustrate this lets go back to our simple A, B, and C network example. When C connects to B, what would motivate B to announce C's networks across to A? Don't forget that in doing this the traffic flowing between A and C will consume B's network resources, but B does not have a direct relationship with either of the end users who are generating this traffic flow. So B cannot bill either A or C's customer to compensate it for providing this service. One viable solution here is for C to pay B for this transit service. In effect, B is C's provider, or, viewed from the other side, C is B's customer. In theory it might be possible to organise the world of interconnected networks into a connectivity mesh using only customer-provider relationships, but at times it might get tricky. To illustrate this, lets add a another network to our example, network D, who is a customer of A. Now when D exchanges traffic with C then D will pay A and C will pay B, as we would expect for customer provider relationships. But what about the relationship between A and B? Sometimes, when A and B are a similar size and scale it is not easy to naturally define who is the provider and who is the customer. The ISP industry has devised an additional form of relationship to address precisely this situation, which is the "peer" relationship, where the two networks interconnect, but agree not to invoice each other (this was the old SKA, or "Sender Keep All" arrangement).

So we have three roles for a network in the domain of interconnection: customer, provider and peer, and many networks have all three relationships at once. Bearing in mind that a network generates revenue from its customers, spends money on its providers and is revenue neutral with its peers, then its clear that providers would like to maximise preference with its customers over peers and providers, and prefer peers over providers. that way a provider can maximise revenue and minimize expenditure.

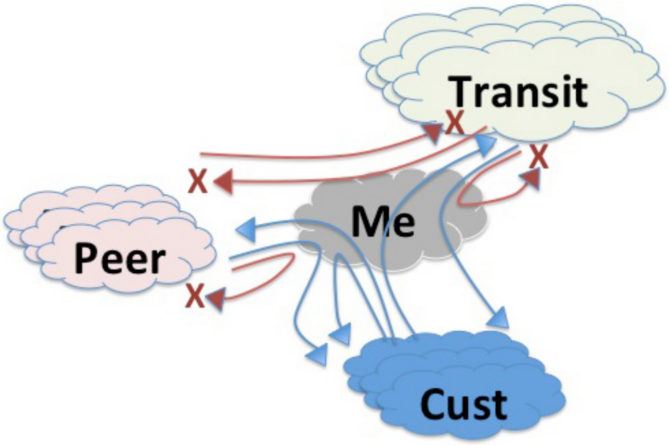
The way this is implemented in a network's routers is by using "local preference" settings in BGP. It all external connections are categorised simply into one of these three categories, then the local preference setting can be used to prefer customer-announced routes over peer-announced routes over provider-announced routes. So if a network sees the same route being advertised from a customer and from a peer or a provider, the local preference setting is intended to ensure that the network will prefer the path via the customer over the path via the peer.

These local preference settings have a high precedence in the BGP decision-making algorithm, and local preference overrides the default BGP comparison algorithm that compares AS path length. So even if a network uses AS path prepending to attempt to bias the path selection, the local preference setting will override this.

It's also helpful to understand re-advertisement preferences in a network, as this too is part of the process of a network attempting to optimise its position by maximising revenue and minimising expenses, and stopping "free riding" where the network is used by unfunded traffic.

To do this most network used the following basic redistribution rules:

- customer-learned routes are re-distributed to customers, peers, and providers
- peer-learned routes are re-distributed to customers but not to other peers nor to providers
- provider-learned routes are re-distributed to customers, but not to other providers, nor to any peers.



So now we have the elements of understanding in a little more details what may have happened between Telstra and Dodo Internet on that Thursday afternoon in Australia.

What?

What happened on Australia on the afternoon of Thursday 23rd February?

One likely explanation here is that some form of accident or misadventure in the Dodo network (AS38285) altered the BGP configuration on an edge router such that it announced its entire internal route set to the Telstra network. As Dodo appears to be a customer of Telstra, then Telstra was prepared to use these routes in preference to routes announced by its peers and providers, as per the preference settings outlined above.

In this case Dodo is a multi-homed network, and not only does the Dodo network contain routes relating to its own customers and services, it is also connected to another transit provider, Optus, (AS7474), and it also contains transit routes from the PIPE Internet Exchange (AS 23745, AS18398) and the Equinex exchange (AS24115).

In looking back out the logs of BGP from the day, the initial sign of a problem was an announcement re-advertised by Telstra (AS1221), received at 02:39 UTC on the 23rd February:

```
        2012/02/23 02:39:45 rcvd UPDATE prefix 23.37.112.0/20,
                                        path   4608 1221 38285 24115 209 20940 20940
```

AS209 is Qwest, a transit network provider in the US, and the normal path to this network is the path:

```
                                        path   4608 1221 4637 209 20940 20940
```

So when a Qwest transit appears in a route from a customer of Tesltra, then this is strong indicator of some form of failure in the routing system. In this case the route is being advertised by Telstra because Telstra has preferred the Dodo route. This is not because the original route is shorter in terms of AS Path length (it's actually longer) but because Telstra International (AS4637) is Telstra's transit provider, and Telstra routing policy is to prefer customer-announced over transit-announced routes, regardless of the AS Path length of the routes in question.

The next 1400 BGP updates from Dodo to Telstra announced the routes from the Pipe Sydney Internet Exchange (AS18398) and the Equinix Echange (AS24115). Then, one minute later Telstra started receiving routes from Dodo that had Optus (AS7474) as its next AS hop:

```
        2012/02/23 02:40:17 rcvd UPDATE prefix 192.142.128.0/24
                                         path  4608 1221 38285 7474 4804
```

This is now a massive problem for Telstra. Optus (AS7474) is a peer of Telstra, and normally Telstra would use this direct peer route to reach only Optus's customers, but, once more, as Dodo is a direct customer, Telstra learns this new route via Dodo and proceeds to prefer it internally and to announce it to all its customers. Optus is also a full service transit provider to its customers, including Dodo, and Optus is announcing all 400,000 Internet routes to Dodo. Dodo now proceeds to announce this entire route set to Telstra, and Telstra prefers this customer path and also proceeds to re-advertise these routes to all of its customers. As this is a customer route it is likely that Telstra is also announcing this to all of its peers although I don't have access to BGP logs that would verify this supposition, and this may cause further disruption. Its also evident that Telstra also announces this route set to Telstra International (AS4637) its transit provider, and this causes further disruption in terms of connectivity to networks that are located in other countries for a selected set of networks that use Telstra International as their transit provider (once more, as Telstra is a customer of Telstra International, Telstra International is preferring routes learned from Telstra, and proceeds to readvertise these false routes to its other customers and its peers.

Now the problem emerges. Now Telstra sees and prefers routes from a customer which Telstra prefers over the routes it receives from its transit providers and peers. And this route set encompasses the entire Internet. At this point Telstra starts directing large amounts of traffic that it would normally pass to its transit provider to its customer Dodo. And not surprisingly at that point things start to fail within Telstra's network, and for all other customers of Telstra, and potentially for some peers of Telstra. The problem spreads to Telstra's main transit provider, Telstra International, and its customers and peers start to experience a subset of the connectivity issues that are being experienced by Telstra itself.

This announcement of routes persists for the next 13 minutes, until 02:53 UTC. At this point it appears that there was some failure in Telstra's network as we see some withdrawals from Telstra and re-advertised routes from a longer transit path that include Telstra International then Telstra and Dodo:

```
        2012/02/23 02:53:45 rcvd UPDATE prefix 93.184.223.0/24
                                        path   4777 2516 4637 1221 38285 7474 7473
                                               3320 15133 15133
```

However, some 60 seconds later the feed of Dodo routes resumed from Telstra.

The routing condition persisted until 03:15 UTC, or a total of 46 minutes for the event.

I should note that while the time of day shown here might indicate that this might have been a middle-of-the night problem, the east coast of Australia is at UTC+11 hours, so this outage started at 1:30 pm on a Thursday afternoon, which is one of the peak business usage periods for Australia. I should also note that Telstra is one of the major transit providers within Australia, so a failure that affects Telstra's customers affects a lot of Australian Internet users. When we include areas of partial failure because of the re-advertisement of these routes by Telstra International then the scope of the outage broadens into one that extends across a number of countries, although to a far lesser extent than that experienced by Australia users.

Why?

Why did we see this failure?

Finger trouble in configuring routers is not a rare event, and potential route leaks happen very often. But most such events come and go without any impact whatsoever. Why was this particular event such a problem? And why aren't other similar events a problem?

Here's where filters come into play. A conventional approach to managing customers, and often peer connections, is to use input routing filters. The input routing filter is intended to specifically limit a customer to announce precisely those routes that the customer has agreed with the provider in advance that it is authorised to announce. If the customer announces further prefixes beyond what is described in the input filter, then of course the filter will remove these extraneous routes before they are learned by the provider.

A plausible, and highly likely explanation of the event here is that there was no input routing filter on the Telstra router, nor any corresponding output routing filter on the Dodo router. Admittedly there are a lot of routes: Dodo originate the equivalent of slightly more than a /11, using a span of 843 separate advertised prefixes. It provides transit to 5 other networks, and announces a further 252 prefixes as a transit to Telstra. This is a total of 1095 separate prefixes, and it is possible that this number is too large for their operational support systems to maintain per-prefix route filters. However, the reason for the lack of filters is speculation. The observed outcomes were consistent with the observation that no route filters were in place at this time of the Dodo to Telstra interconnect.

When the Dodo route configuration changed, then the internal routes used by Dodo leaked to Telstra, and Telstra inappropriately learned the entire Internet via a path through Dodo, and then proceeded to redirect all its egress traffic to Dodo, with the consequent, and widely reported, disruptions to service. It also passed these routes to its other customers, to its peers and to its transit providers, and the disruption spread out across more networks.

The Fix?

It would be nice to think that we could fix this, and do a better job in running a distributed routing system that does not leak and fail from time to time, but all the evidence suggests that these leaks have been an intermittent "feature" of the Internet for about as long as the Internet itself. A record of detected route leaks in the period from 2003-2010 can be founds at the URL: http://dyadis.cs.arizona.edu/projects/lsrl-events-from-2003-to-2009. The authors argue that they detected between 5 to 20 large scale routing leaks per year over that 6 year period.

The conventional approach to preventing route leaks is to maintain route filters.

Route Filters?

A network operator can insist that all customers and all peers enumerate specifically the list of prefixes that they intend to announce. The network operator can use these lists to maintain filter lists on the edge routers to the network's customers and peers. When a route is received, the route can be passed through the filter list, and is only accepted once it passes through the filter.

For customers and peers that present with a small number of prefixes this can be maintained relatively easily, but its an approach that does not have good scaling properties. Its one thing to run filter lists for a handful of customers, each with a handful of routes, but when the numbers start to head into the hundreds, then its a case of using automated tools. And when the numbers rise again into the thousands then the efforts of maintaining large filters, even with various operational support tools starts to get quite challenging.

Maintaining route filters for the larger providers pose operational challenges in terms of escalating administrative overhead, cost of maintenance, and accuracy and timeliness of the entries that are in the filters. For a customer while there may be a strong motivation to add new entries to the filter on a timely basis, there is actually no motivation to remove the out-of-date entries, and the consequent filter bloat becomes a real challenge to manage.

It seems that when the collection of routes gets sufficiently large, or when then the level of administrative updates in terms of adds, removals and amendments gets too large, then providers often choose to take each other "on trust" and drop the use of administratively maintained routing filters.

At this point filtering based on AS path rather than by prefix starts to look tempting. It is possible to augment, or even replace, the filter lists of prefixes with filter lists of AS Paths. In this case if the other party attempts to re-advertise learned routes, then the AS Path of these routes would trigger the filter action.

Unfortunately its not as good as it sounds. Lets look at why in the case of Telstra and Dodo once more.

Dodo announces 843 prefixes that are originated by AS 38285 (Dodo). However, in the "normal" state it also announces 252 routes for third parties. These third party ASes are transits for others ASes and so on. So the first problem is that even though it may be possible to limit the AS paths accepted from a BGP neighbour, this "limit" may still encompass potentially large swathes of the Internet, and a damaging route leak may still occur even within the parameters of AS Path filters.

It's also the case that not all route leaks are as "well" behaved as this incident. This route leak was an example of an unintentional re-advertisement. Other forms of route leaks have involved mapping externally-learned eBGP routes into the IGP and then mapping all IGP routes back into BGP and passing them out to the peer as if they were originated directly in the network. Another form of route leak involves leaking out a bevy of more specific internal routes to externally connected networks. In this latter case there is no direct subversion of third party routes, but if the internal route set encompassed a million or more routes, the leak of such a large volume of routes into the inter-domain routing space would likely trigger a number of limits and result in BGP session teardowns and consequent third party connectivity damage. In both of these cases the advertised leak looks like the local AS is the originator, and an AS Path filter would not be effective in managing the leak.

Is filtering the only approach? This thought then leads to some further questions: Can we do a better job without necessarily involving manually-maintained filters? Indeed, can we go one better and devise routing control systems that detect and react to route leaks and supress them automatically?

RPKI and Route Leaks?

In thinking about this, it's also useful to bear in mind that many potential security problems are exposed by accident, and while it may at first be misadventure that exposes a vulnerability, thereafter the vulnerability is deliberately manipulated for adverse reasons. So in this case its not unreasonable to turn to the current efforts to secure BGP (http://www.potaroo.net/ispcol/2011-07/bgpsec.html), and ask the question: Can the RPKI fix this? If this secure routing framework is capable of detecting and filtering out BGP routes that have been tampered with or falsified, then surely secure BGP would allow us to identify and filter out route leaks. Yes?

However this is not going to happen, or at least not with just the secure BGP protocols that are currently being developed in the context of the ITEF's SIDR Working Group.

Lets have a quick look at this, as it is not the most obvious of answers.

If you are receiving a route with an AS path of the form <A B C>, and the origination of the prefix at C is verified, then the only way you can identify that this route is an unintentional leak, as compared to the conventional operation of BGP, is not by looking at the operation of protocol per se, but by looking at the routing policy intentions of A, B and C, and working out if what you are seeing with the AS Path <A B C> is intentional within the scope of the routing policies of these entities. But secure BGP does not contain routing policy information. Secure BGP can allow you to verify that the holder of the prefix authorised C to originate a route, which it is doing. Path security in secure BGP can also allow you to verify that C passed the advertisement to B, who, in turn, passed it to C. So from the perspective of secure BGP there is nothing invalid about this route, and cannot inform you whether it is an intentional advertisement of a route or an unauthorised route leak.

It exposes a broader issue here about the difference between routing intent and routing protocol operational correctness. A protocol correctness tool, such as secure BGP, is able to tell you that the routing information has been faithfully propagated across the network via the operation of the routing protocol, but such a tool cannot tell you whether the routes that are being propagated were intentionally distributed or not.

How could we do this?

Route Registries?

The use of Internet Routing Registries and the associated Routing Policy Specification Language (RPSL) (RFC 2622, RFC40122) is an alternative approach to the manual management of route filters. RPSL is a relatively rich language and, as the name says, it allows a user to describe a network's import and export policies in terms of relationship with adjacent AS's and its transit (re-advertisement) policies.

If this is used in the context of a routing registry it allows a network operator to enumerate the prefixes originated by the local AS and the transit policies that are associated with these routes. It also allows the network operator to describe its re-advertisement policies by specifying its AS neighbours and the routing policies applied to routes learned from adjacent ASes.

If every AS maintained an accurate, up-to-date and complete set of prefix and route policy entries in an Internet Routing Registry, then it appears that it would be theoretically possible for an AS to generate a prefix and AS path filter set for all of its network adjacencies through a computation across the registry's contents. Indeed there are tools that attempt to do precisely that for the existing route registries.

Why aren't we all doing precisely this? Why aren't we using these route registry tools as part of our standard operating practice?

The story about the use of route registries is a very mixed one.

They have been around for almost twenty years now in one form or another, and some regions of the world have been very diligent in compelling every network operator in their region to maintain accurate information in their local routing registry. But in other cases the route registry story is not so encouraging.

RPSL is a complex language and it can be challenging to accurately describe the intricacy of some routing policies in RPSL. Its often the case that the registry is populated with "just in case" entries, as well as historic entries, so sorting out what is current routing intention from other extraneous data in the registry is extremely difficult, and to do so with an automated registry scanning tool has proved not to be possible so far. Its also the case that network operators often use a level of granularity of each eBGP session between adjacent ASes, while RPSL uses a coarser level of granularity of individual ASes. It is therefore more challenging to describe the individual routing policies that apply to each BGP session between the same two ASes, and there is also the question as to whether network operators would be comfortable in publishing such a detailed level of information about their network's routing policies.

The route registries we use today have various models of authenticity and integrity. It's possible in many cases for a registry user to enter routing information for third party prefixes without the authority of the actual prefix holder. Sorting out what is recognisable as authoritative information from what is not authoritative is not helped by a registry data model that typically includes no validation or authority information. There are also many route registries, and its often the case that they contain conflicting information. Which registry should be "preferred" if one wanted to resolve these contradictions in information? Why?

> A NANOG presentation from October 2008 is still once of the better summaries of the problems we face with route registries. I do not believe that much has change in the 4 years since this presentation was given:
> http://www.nanog.org/meetings/nanog44/presentations/Tuesday/RAS_irrdata_N44.pdf

This would be challenging enough, but the problem is further compounded by the observation that in many areas of the Internet operators have eschewed the route registry approach and rely on their own customised tools. So not only is the quality of the information in route registries variable, the coverage of the information in route registries is also variable.

Where to from here?

Some longstanding problems are longstanding because we have not quite managed to apply the appropriate analytical approach to the problem. There is a solution out there, but it involves some searching!

We could try, yet again, to coerce the industry to diligently use route registries for all external routing, but what would be different from this call to use route registries from all the other calls in the past? And if its no different, then why would such a call enjoy any greater levels of take up than has happened in the past?

Maybe we could use digital signatures and the RPKI to combine information authenticity with the route registries. However the issue we may want to consider in this case is would this only make an already complex and difficult system yet more complex and even harder to use?

Maybe this particular problem is a different kind of longstanding problem. Some problems are longstanding problems simply because they are just exceptionally hard problems!

This makes me wonder if there are alternate perspectives on the space we are working in. For example, would we think about this problem differently if we were to think about routing not as a topology and reachability tool, but an distributed algorithm to solve a set of simultaneous equations. The equations here are expressions of routing policies, and the aim of the algorithm is to converge on solutions that solve individual equations as well as converging on a network-wide solution of maximal connectivity. Would such a perspective provide a different insight as to the way in which routing policies and routing protocols interact? And could such a perspective provide some leads as to how we could not only secure the routing system against deliberate abuse and malfeasance but also secure it against inadvertent misadventure in the form of route leaks?

Disclaimer

The views expressed are the author's and not those of APNIC, unless APNIC is specifically identified as the author of the communication. APNIC will not be legally responsible in contract, tort or otherwise for any statement made in this publication.

About the Author

Geoff Huston B.Sc., M.Sc., has been closely involved with the development of the Internet for many years, particularly within Australia, where he was responsible for the initial build of the Internet within the Australian academic and research sector. He is author of a number of Internet-related books, and has been active in the Internet Engineering Task Force for many years.

www.potaroo.net