Poster: IRR Hygiene in the RPKI Era
Ben Du
UC San Diego
bendu@ucsd.edu
Alex C. Snoeren
UC San Diego
snoeren@cs.ucsd.edu
kc claffy
UC San Diego
kc@caida.org

Problem and motivation
The Border Gateway Protocol (BGP) is the protocol that networks use to exchange (announce) routing information across the Internet. Unfortunately, BGP has no mechanism to prevent unauthorized announcement of network addresses, also known as prefix hijacks. Since the 1990s, the primary means of protecting against unauthorized origin announcements has been the use of routing information databases, so that networks can verify prefix origin information they receive from their neighbors in BGP messages. In the 1990s, operators deployed databases now collectively known as the Internet Routing Registry (IRR), which depend on voluntary (although sometimes contractually required) contribution of routing information without strict (or sometimes any) validation. Coverage, accuracy, and use of these databases remains inconsistent across ISPs and over time.

In 2012, after years of debate over approaches to improving routing security, the operator community deployed an alternative known as the Resource Public Key Infrastructure (RPKI). The RPKI includes cryptographic attestation of records, including expiration dates, with each Regional Internet Registry (IRR), which depend on voluntary (although sometimes contractually required) contribution of routing information without strict (or sometimes any) validation. Coverage, accuracy, and use of these databases remains inconsistent across ISPs and over time.

Although RPKI use is growing, its limited coverage means that security-conscious operators may query both IRR and RPKI databases to maximize routing security. However, IRR information may be inaccurate due to improper hygiene, such as not updating the origin information after changes in routing policy or prefix ownership. Since RPKI uses a stricter registration and validation process, we use it as a baseline against which to compare the trends in accuracy and coverage of IRR data.

Related work
Researchers have compared prefix-origin pairs in both IRR and RPKI to those observed in BGP announcements and routing tables, finding 80% (2013) [2] and 90.4% (2019) [1] of records are consistent with BGP announcements, respectively. However, we know of no detailed comparisons between IRR and RPKI databases since the launch of RPKI.

Dataset and methodology
Our dataset consists of the IRR dataset, a longitudinal dataset of an IRR database, and RPKI dataset, a longitudinal dataset of RPKI databases. To build the IRR dataset, we collected the IRR database archives from the Routing Assets Database (RADB). RADB provides the largest IRR database mirror on the Internet, including IRR databases hosted by 34 other organizations such as NTT and RIPE NCC. We downloaded monthly snapshots of RADB from August 2016 to September 2021 and extracted route objects for their IP prefixes and origin AS information.

To build the RPKI dataset, we collected validated ROA objects (verified prefix origin information) from RIPE NCC’s RPKI validator. RIPE NCC publishes validated ROA objects from all five RPKI trust anchors (APNIC, ARIN, RIPE NCC, AFRINIC, LACNIC). We download the monthly ROA archive starting August 2016 to September 2021 to match the interval of our IRR dataset.

Our methodology explores differences between the information registered in the IRR dataset (IRR) and the RPKI dataset (RPKI) from the following two aspects: database completeness and record consistency. To study database completeness, we look in each dataset at the number of IP prefixes and ASes and their coverage of the allocated IPv4 address space. To study record consistency, we take all records in our IRR dataset and perform RPKI validation on those prefixes, similar to the mechanism of BGP route origin validation. We put the IRR records into 4 categories – consistent, inconsistent ASN, inconsistent maxLength, and not in RPKI by applying the following validation logic:

1. For each record $R_x$ in the IRR dataset, we denote the prefix as $P_x$ and origin AS as $AS_x$.
2. We look for an exact matching prefix or covering prefixes of $P_x$ in the ROA objects in the RPKI dataset. The resulting list of candidate ROAs are denoted $L_ROA$.
3. If $L_ROA$ is empty, then we put $R_x$ in not in RPKI.
4. For each candidate ROA, $C_ROA$, in $L_ROA$, we put $C_ROA$ in a list, $M_ROA$, if the origin AS in $C_ROA$ equals $AS_x$.
5. If $M_ROA$ is empty, then we classify $R_x$ as inconsistent ASN.
6. For each $C_ROA$ in $M_ROA$, we put $C_ROA$ in a final list, $V_ROA$, if the prefix length of $P_x$ does not exceed maxLength field in $C_ROA$. 
7 If $V_{ROA}$ is empty, we classify $R_x$ as inconsistent maxLength, otherwise as consistent.

Results and contribution

Our results summarize the completeness and consistency between the IRR dataset and RPKI dataset. Table 1 shows that IRR dataset contained 38 times more prefixes and covers almost 7 times more allocated IPv4 address space than RPKI in 2016. Table 1 also shows that RPKI almost doubled its number of prefixes over 6 years, more rapid growth than IRR.

<table>
<thead>
<tr>
<th>Year</th>
<th>Prefix</th>
<th>IRR ASN</th>
<th>IP Space</th>
<th>Prefix</th>
<th>RPKI ASN</th>
<th>IP Space</th>
</tr>
</thead>
<tbody>
<tr>
<td>2016</td>
<td>769k</td>
<td>24,112</td>
<td>70.52%</td>
<td>20k</td>
<td>3,741</td>
<td>11.62%</td>
</tr>
<tr>
<td>2017</td>
<td>813k</td>
<td>27,151</td>
<td>73.39%</td>
<td>34k</td>
<td>4,918</td>
<td>14.25%</td>
</tr>
<tr>
<td>2018</td>
<td>900k</td>
<td>30,531</td>
<td>74.23%</td>
<td>44k</td>
<td>6,185</td>
<td>15.08%</td>
</tr>
<tr>
<td>2019</td>
<td>958k</td>
<td>33,608</td>
<td>74.73%</td>
<td>75k</td>
<td>9,349</td>
<td>23.55%</td>
</tr>
<tr>
<td>2020</td>
<td>1M</td>
<td>37,427</td>
<td>82.59%</td>
<td>128k</td>
<td>15,039</td>
<td>35.06%</td>
</tr>
<tr>
<td>2021</td>
<td>1.06M</td>
<td>40,574</td>
<td>92.73%</td>
<td>209k</td>
<td>23,472</td>
<td>49.26%</td>
</tr>
</tbody>
</table>

Table 1: RPKI is growing faster than IRR, but the IRR dataset is still more complete than the RPKI dataset.

<table>
<thead>
<tr>
<th></th>
<th>AS Class</th>
<th>AS Count</th>
<th>Consistent</th>
<th>Inconsistent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entirely Consistent</td>
<td>4326</td>
<td>31,897</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Entirely Inconsistent</td>
<td>3600</td>
<td>0</td>
<td>47,395</td>
<td></td>
</tr>
<tr>
<td>Mixed</td>
<td>2040</td>
<td>123,609</td>
<td>156,552</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: More ASes keep their entire IRR records consistent with RPKI.

Figure 1: There are more conflicting records than agreeing records between IRR and RPKI.

Record consistency We observed substantial inconsistency between IRR records and corresponding ROAs in RPKI. The number of IRR records in each category has grown as RPKI gained popularity, but notably there are more inconsistent records between IRR and RPKI than consistent ones (Fig. 1). The purple line reflects a significant uptick of inconsistent ASN records in October 2016, when 26 thousand more IRR records were inconsistent with their corresponding RPKI records. Those 26 thousand IRR records were registered under Verisign’s AS number, with a description of verisign customer route. In October 2016, those customer ASes of Verisign registered their prefixes in RPKI under their own AS numbers, causing significant inconsistency between IRR and RPKI. In September 2019, Verisign deleted those conflicting records from the IRR, causing the purple line downtick in Figure 1. In January 2019, the green line shows an increase of consistent records, caused by TWNIC ASes bulk registering their prefixes in RPKI.

The main contribution of this work is to explore IRR hygiene by comparing completeness and consistency between IRR and RPKI. We find the rapid growth of RPKI adoption helpful for measuring IRR correctness and better routing security. However, because lack of consistency suggests stale IRR data, tools that identify such inconsistencies can help those wanting (or willing) to maximize the utility of both platforms. We will expand our analysis to correlate usage and consistency with other network properties, e.g., network size, location, type, that reveal insight into how the ecosystem is evolving.

REFERENCES