A solution to the problem with IPs

UKSG eNews 381

30 Sep 2016

Andrew Pitts, Publisher Solutions International

PSI presented two workshops at the UKSG annual meeting in Bournemouth earlier this year. They were extremely well attended with people having to stand on both days. There were many questions and the workshops exceeded their allocated time slots. My co-presenters were Keith Abbott from Wiley and Charlie White from SAGE. We demonstrated findings from a three year project to collect, aggregate and verify IP addresses from our publisher partners.

We have built a strong relationship with publishers due to our work to stop the global spread of subscription abuse by intermediaries. Over the last twelve years, we have monitored data from dozens of publishers and have carried out circa 220 investigations, winning more than $40 million USD in damages for the publishing industry. We found everything from personal rate subscription abuse, username and password abuse, Athens IDs being sold and, of course, IP address abuse. Our investigations have taken us everywhere from the USA, South America, Europe, Asia and Africa; IP abuse is prevalent everywhere.

We first started looking at IP addresses because we found instances where publishers were not receiving the full amounts paid by institutions because multiple libraries were included in one licence and the IP addresses folded up into one range; such abuse was previously impossible to verify. Millions had been lost to the publishers and they and the institutions were completely unaware this was happening. We decided to discover the extent to which this was happening and what publishers and institutions could do about it. We partnered with Wiley and with the help of Roy Kaufman, their assistant General Counsel, we looked into the causes and possible prevention. Wiley had a team of six manually type each IP address held for customers in Poland into Who Is searches to see how easy it would be to verify each one. This one country took the team three weeks and it found that in 66% of searches, the Internet Assigned Numbers Authority (IANA) record merely advised that the IP address belonged to an internet service provider (ISP) eg British Telecom or Comcast. This was alarming as even found institutions were often incorrect but the exercise clearly demonstrated how easy it was to abuse IP addresses.

Next we asked IANA why the registry was incomplete and why so many records were attributed to ISPs. We visited RIPE, the European IP Numbers Authority, and asked how we might find the exact owners of the IP addresses and were advised that IPs were never intended to be a record of ownership and that there was no onus on ISPs to inform the registry to whom IPs had been assigned. We were, however, after receiving letters of recommendation from several of the world's largest publishers, offered a download of the entire registry to help with our efforts when required.

We tried looking up domains but realised this too was ineffective because institutions do not give all their owned IP ranges to publishers for verification purposes, and vital ranges were missing from

third parties like OCLC's EZ Proxy service, ProQuest, EBSCO and a host of others. We tried sending thousands of letters to institutions explaining what we were doing and also had publishers writing to their customers asking for up to date IP addresses. This also proved futile with only a few institutions replying. From over 5,000 emails in Mandarin with endorsements from publishers sent to known Chinese libraries, we only received a handful of responses. Imagine trying in dozens of different languages to get over 60,000 institutions worldwide to send you data!

We went back to the drawing board and approached the problem in an entirely different way. With our trusted reputation for gathering and holding sensitive data (being in Europe really helped here, with our extremely strict laws on data management), we received customer data from over 150 STM publishers of all sizes and types. This allowed us to carry out a global clean-up of IP addresses and create a database of verified ranges. It took a few weeks to develop the logic but we ended up, country by country, being able to ask 27 different questions of the data to come up with reports to publishers showing the correct and incorrect IP addresses for their existing customer base. We analysed data and IP addresses from 218 countries, territories, or regions and created a database containing naming standards for over 60,000 academic, corporate and government institutions. This amounted to a staggering collection of IP addresses totalling more than 1.5 billion which grows on a daily basis.

The project proved that it was indeed impossible to search with any accuracy, in fact a 33% accuracy, the IANA data. It also showed that 58% of the collected IP addresses required investigation because of inaccuracies caused by the following;

- mistyped IP addresses opening access to unintended institutions with the mistyped IP never being discovered and deleted: manually typing large strings of numbers at so many points in the supply chain had caused a plethora of errors, often perpetuated over and over again if not spotted
- a great reluctance by publishers to delete IP addresses at all when asked for fear of disrupting a library's access, resulting in huge IP address duplication and overlapping
- ranges which are far too large being applied by accident: we saw one range that was overblown to include 40,960,000 IP addresses opening access to over 350 institutions in 42 different countries
- usage stats being incorrect as duplicated IP addresses, large ranges and incorrect IP address allocation is not captured in COUNTER usage statistics: it is all well and good having double clicks removed but what about when you are receiving the stats from several institutions because you had too large a range applied to your account
- mistyped IPs by libraries or agents which can result in multiple publishers all receiving the incorrect IP addresses
- not everyone being as trustworthy as they could have been, with multiple institutions being provided to publishers to avoid extra licence costs

In some instances, we discovered that agents had been adding their own IP address to those of institutions provided to publishers over many years giving them unfettered access to collections. Databases were created and sold on to institutions, papers were printed and provided to libraries and licences were manipulated in order to pass off several institutions as one licence.

Some of these issues were highlighted by Keith and Charlie in the workshop.

Keith explained that "12 separate institutions in Germany share an IP range, for example 111.222.*.*. Entering that single IP address will authenticate everyone in those 12 institutions whether they have paid for the content or not. It will also attribute a download to each organisation the publisher does have a deal with - thereby incorrect usage statistics are recorded. Wiley has been working with PSI for 11 years to rectify the problems incorrect IP data brings".

Charlie showed first-hand what it is like from the publisher perspective when IP addresses are misused, citing an example of a situation that dozens of publishers faced when an agent in Thailand

decided to manipulate IP addresses. In turn, this affected the usage stats for every institution involved.

The three of us believe the industry needs a practical, economically viable and effective solution for managing IP addresses and that is why we have created the IP Registry and plan to make the database available to publishers and libraries globally early in 2017.

This UKSG Editorial is taken from the industry newsletter *UKSG eNews,* published every two weeks exclusively for UKSG members. The newsletter provides up-to-the-minute news of current issues and developments within the global knowledge community.

To enjoy *UKSG eNews* and other member benefits become a UKSG member. To submit an editorial suggestion for *UKSG eNews,* contact the editors: seneditor@uksg.org.