A Structural Model of Internet Organization Discovery

Zi-yu Yang, Xiao-yun Wang, Hong-mei Ma and Li Qin
Library, China Defense Science and Technology Information Center, Beijing, China

Keywords: Organization Discovery, Social Network, IP Allocations, Routing Registry.

Abstract: This paper presents a highly structured model to automatically discover Internet organizations from the data of RIR (Regional Internet Registry) and IRR (Internet Routing Registry), where network operators register their networking resources such as IP addresses and routing policies. Our basic idea is to discover network operators that have close ties among each other from those registry activities, and consider them as being from the same organization. With the data from two RIRs, this model produces to date the first organization level network of current Internet. The model shows its reasonability in our validation with real Internet routing data, and is likely to be applied extensively in networking area.

1 INTRODUCTION

The Internet is running under the administration of thousands of organizations, each of which can be an institution, a company or a university. An *Internet Organization* (*IORG*) is an organization that uses at least an IP network to host application services or access the Internet. Despite the important role of *IORG*s in advancing the Internet forward, neither those organizations nor the relationships in between have been well understood or characterized yet. In fact, even how to identify *IORG*s is still an open question, making the further studies such as their formed ecosystem impossible.

IORG discovery is difficult for several reasons. First of all, considering the tremendous size of the Internet, the amount of organizations is huge as well. Secondly, there is no authoritative source of *IORG* data. Last but not least, as the growth of the Internet, *IORG*s are highly dynamic. Some *IORGs* may vanish while some new ones may appear. Therefore, any approach that used to discover *IORGs* should be highly structured and automated such that the discovery process can be regularly repeated with new input.

In this paper, we design and propose a highly structured model to automatically discover *IORG*s from the data of RIR (Regional Internet Registry), where network operators register their networking resources such as IP addresses and routing policy. Let network operators be the representative of organizations in the registry activities, our basic idea is to discover network operators that have close ties among each other from those registry activities, and consider them as being from the same organization.

To our best knowledge, this has been no similar research in the direction of *IORG* discovery so far. Researches closest to ours are (Siganos and Faloutsos, 2007) and (Cai and Heidemann, 2010). Siganos *et al*. used the allocation records from RIRs, registered ISP (Internet Service Provider) routing policy from IRR(Internet Routing Registry) to detect erroneous and suspicious routing behaviour. However, they restricted their work on data only, and did nothing about *IORG*s. Cai *et al*. aimed for an *AS*(Autonomous System)-*to-Organization* map that allows a more accurate view of Internet in the granularity of AS. Since the size of organizations vary largely, and only large organizations are qualified to apply for AS numbers, their work actually focused on only large organizations. On the contrary, we provide a finer granularity to observe the organization-level Internet. Beside, different from that both these two studies rely on ad-hoc methods and require a large amount of manual intervention, our model is highly structured and automated.

The remainder of this paper is structured as follows. Section 2 introduces the registry activities of network operators that serve as the base of our model. We present the discovery methodology in section 3. The discovery and validation results are shown in section 4, and we conclude this paper in section 5.

Yang, Z., Wang, X., Ma, H. and Qin, L..
A Structural Model of Internet Organization Discovery.
In *Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015) - Volume 1: KDIR*, pages 344-349
ISBN: 978-989-758-158-8

2 INTERNET REGISTRY ACTIVITIES

The virtual activities we consider to be the mirror of *IORG*s activities in real world are Internet registry activities, which can be divided into two categories, regarding Internet number resources and routing policies respectively.

2.1 Registry of Resource Allocations

Internet number resources refer to IP addresses and AS numbers, which are essentials to access the Internet or perform network management on AS level.

IANA (Internet Assigned Numbers Authority) serves as the root of resource allocation chain to ensure the unique use of Internet number resources. RIRs are set up to coordinate the allocation of those resources inside their own regions on behalf of IANA. There are currently five RIRs: ARIN, APNIC, AFRNIC, RIPE and LACNIC. RIRs subsequently allocate number resources received from IANA to NIRs (National Internet Registry, e.g. CNNIC), or directly to LIRs (Local Internet Registry, e.g. AT&T) which are usually large ISPs (Internet Service Provider). NIRs and LIRs can further allocate the resources they received from RIRs to end users or other ISPs.

The allocation activities of number resources are required to be registered in related RIR databases. To ensure the completeness and freshness of registry information, RIRs usually carry out strict policy to assure that allocation activities are registered in time.

```
inetnum:      62.0.0.0 - 62.255.255.255
netname:      EU-ZZ-62
descr:        RIPE NCC
descr:        European Regional Registry
country:      EU
org:          ORG-NCC1-RIPE
admin-c:      CREW-RIPE
tech-c:       CREW-RIPE
tech-c:       OPS4-RIPE
status:       ALLOCATED UNSPECIFIED
mnt-by:       RIPE-NCC-HM-MNT
mnt-lower:    RIPE-NCC-HM-MNT
changed:      hostmaster@ripe.net 19970428
changed:      hostmaster@ripe.net 20020408
changed:      ripe-dbm@ripe.net 20040422
source:       RIPE

inetnum:      62.0.4.0 - 62.0.4.255
netname:      NV-GILAT-VSAT
descr:        NV-GILAT-VSAT
country:      IL
admin-c:      AL1028-RIPE
tech-c:       NN105-RIPE
status:       ASSIGNED PA
mnt-by:       NV-MNT-RIPE
mnt-lower:    NV-MNT-RIPE
changed:      noc-team@netvision.net.il 20041130
source:       RIPE
```

Figure 1: IP allocations registered by two institutes.

For example, ARIN claims that every allocation or assignment that contains eight or more IP addresses should be recorded in its database. If not, future allocations would be impacted. Similarly, RIPE will check the correctness of relevant registry information when a LIR or ISP requests for a new allocation.

In Figure 1, there are two allocation records of IP address, registered by RIPE-NCC and Netvision company respectively. Based on these two records, we can know that RIPE-NCC has the administration authority over IP addresses ranging from 62.0.0.0 to 62.255.255.255, while Netvision company owns the network 62.0.4.0~62.0.4.255.

2.2 Registry of Routing Policy

Routing registry is used to improve the Internet wide routing by sharing routing policies among ISPs, and the institution in charge is IRR (Internet Routing Registry). An ISP can leverage IRR to publish its routing policy, or look up peering agreements to optimize its routing policy. IRR consists of several distributed databases that usually mirror each other. For instance, RIPE has mirrored more than 10 partners, including ARIN-RR, APNIC, NTTCOM DB, Merit RADB and so on.

```
route:        202.36.121.0/24
descr:        Internet ProLink NZ Limited
descr:        PO Box 91235
descr:        Auckland
descr:        New Zealand
country:      NZ
origin:       AS6831
notify:       support@iprolink.co.nz
mnt-by:       MAINT-IPROLINK
changed:      craig@iprolink.co.nz 981013
source:       APNIC
```

Figure 2: A route object registered by an Internet company.

Figure 2 depicts a route record registered by Internet ProLink NZ Limited, a company located in Auckland, New Zealand. By registering this record in the database of APNIC, this company claims that AS6831 is authorized to originate the network prefix 202.36.121.0/24 in Internet routing system, and ISPs worldwide can use this information to filter false announcements regarding this network.

2.3 Route Policy Specific Language

RIRs use different languages to describe resource allocations and registered routing polices. Currently, ARIN uses SWIP (Shared Whois Project), RIPE and APNIC use RPSL(Route Policy Specific Language), while LACNIC uses a mix of both. Our model is

based on RPSL for its wider usage, and it can be applied to SWIP with tiny modifications.

RPSL is designed to specify routing policy at various levels, ranging from router to AS. In an ideal case, low-level router configurations can be directly generated from the routing policies described at AS level. Like typical object-orientated language, RPSL comprises several classes, each of which uses a set of attributes to describe its object instances. In our model, RPSL classes are classified into three categories: *PoC*(Point of Contact), *NR*(Number Resource) and *RP*(Routing Policy), according to the content being described.

(1) *PoC* classes.

PoC classes describe contact information. For details, *PoC* classes include *mntner*, *person* and *role* class. The *mntner* class specifies authentication information required to add, delete or modify other objects. The *person* class describes the information necessary to contact a person. The *role* class is very similar to person class except for that instead of describing a human being, a role object describes a role performed by one or more human beings. In this way, role does not have to change when a person performing this role changes.

(2) *NR* classes.

NR classes describe Internet number resources, such as *inetnum*, *inet6num* and *domain* class in RPSL.

(3) *RP* classes.

RP classes are used to describe routing policy. For example, the *inet-rtr* class defines a router via this router's DNS name, the IP address of each interface, the AS number of the AS which owns or operates this router and such information.

The *NR* and *RP* objects can establish direct connections with network operators by referring to the class key of *PoC* objects through their *admin-c*, *tech-c* and *mnt-** attributes(including *mnt-by*, *mnt-lower*, *mnt-routes* and so on), as shown in Figure 3.

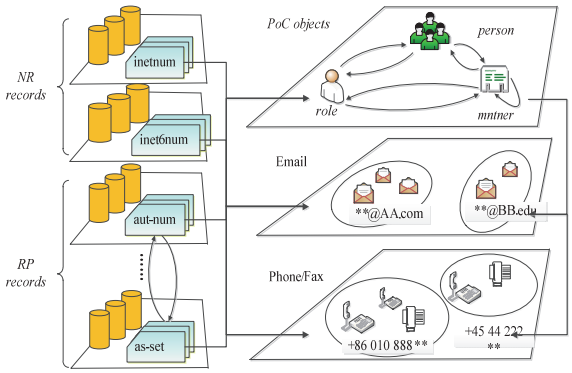

Figure 3: PoC objects are referred as contact points.

While *admin-c* attribute usually refers to someone who is physically located at the site of the network, the *tech-c* attribute indicates a person responsible for the day-to-day operation of the network, but does not need to be physically located at the site of the network.

3 METHODOLOGY

Our methodology is as follows. Firstly, we build a MDN (multiple dimension network) to characterize the interrelationship of various elements in registry data, which is the outcome of network operators' registry activities. Secondly, we quantify how close are two network operators with the tie strength in between, which is calculated based on the paths between these two operators in the built MDN. At last, network operators are grouped into clusters according to the tie strength among them, and each cluster is considered as an organization.

3.1 Building Multi-Dimension Network

Let symbol D={*as-block*,*as-set*,*aut-num*,*inet6num*, *inetnum* ,*mntner*…, *mail*, *phone/fax* number} denote the dimension vector of the MDN, each element in D represents either a RPSL or a user-defined class.

3.1.1 Vertexes of MDN

Let $V_{(i)}$ denote the set of vertexes from dimension $i \in D$, then $V_{(i)}$ should be the union set of all the object instances' class keys of class i. Similarly, the vertex sets of *email* and *phone*/*fax* number dimension are all the email addresses and phone/fax numbers that appeared in the dataset.

3.1.2 Discovering Links from RPSL Objects

MDN links are primarily generated from RPSL objects, each of which is essentially a collection of attributes. For each RPSL object r, it has a key attribute ($r.k$) and a set of non-key attributes ($r.NK$). For each attribute $x \in NK$, the non-key attribute x can be:

(1) *Key of other RPSL objects*. The definition of r leverages the information that has already been defined by some other RPSL objects. In this case, a link $k \rightarrow x$ is added to the link set.

(2) *Plain text*. Since natural language processing is not so accurate, no links are generated in this case to avoid importing uncertainty.

(3) *Email*, *phone*/*fax number*. By referring to email addresses, phone or fax numbers, we know how to reach the personnel responsible for record *r*. In this case, we generate a link between *k* and *x*, and add it to the link set.

3.1.3 Discovering Links via Correlation within Dimension

Within each dimension $i \in D$, our purpose is to discover any two vertexes whose key attributes are related, and generate a link between these two vertexes.

Take dimension *inetnum* as an example, as shown in Figure 1, since the IP address block 62.0.4.0-62.0.4.255 is a subset of another, 62.0.0.0-62.255.255.255, we add a link between these two *inetnum* objects, as shown in Figure 4. This method also works for dimension *as-block*, and we do not repeat here for the sake of brevity.

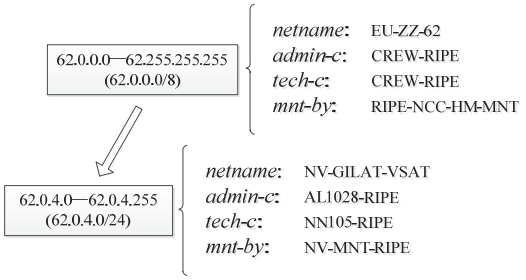

Figure 4: Adding a link between two *inetnum*s relevant.

As for *email* dimension, we manually build a blacklist of domain names that are from RIRs or public email services, such as Gmail, Yahoo, then add a link for every two email addresses that sharing the same domain name.

Correlations within *fax* and *phone* dimension may also be useful in link generation. However, since there is no universe method to parse phone number into country code, region code and institution code, we prefer to be conservative and do not perform this correlation.

3.1.4 Discovering Links via Correlation across Dimensions

Correlations across dimensions occur between set object and its member objects. For example, an *as-block* object usually describes several consecutive AS numbers, thus we can add a link between this *as-block* object and every *aut-num* object whose AS number is included. This theory also works for *irt-set* object (a set of routers) and *inet-irt* object (a router), *rtr-set* (a set of routes) and *route* object(a route), and so on.

3.2 Calculating Tie Strength

We then derive a weighted graph of network operators $G_o=<V_o, E_o>$ that V_o consists of all the *PoC* objects and E_o is the link set. For any two vertexes u, $v \in V_o$, there is a link $(u, v) \in E_o$ if and only if they are connected in MDN. Let $w_{(u, v)}$ denote the strength of link (u, v) in G_o, it is defined to be the accumulated strength of the multiple paths between u and v that may traverse through one or multiple dimensions in MDN.

Moreover, the strength of a path is linearly proportional to the strength of each single link on that path for these links are in series.

Let S be a n-dimension matrix(n is the number of dimensions in MDN that $n=|D|$), and each of its element $s_{i,j}$ ($1 \le i, j \le n$) be the strength of a link between two vertices within the i^{th} dimension and j^{th} dimension respectively, the tie strength of link $(u,v) \in E$ should be

$$w_{(u,v)} = \sum_{l \in P(u,v)} k^{\alpha} \left(\prod_{(i,j) \in l} s_{d_i, d_j} \right)^{1/k} (-1 < \alpha < 0)$$

Where $P(u,v)$ denotes all the paths between vertex u and v, k is the length of path l, a is a constant between 0 and 1, and d_i is the dimension where vertex i is from.

3.3 Classifying Network Operators into Clusters

To group network operators into clusters, we adopted the algorithm presented in (Blondel and Guillaume, 2008), which aims to maximum the network's modularity, defined as

$$Q = \frac{1}{2m} \sum_{i,j} [A_{i,j} - \frac{k_i k_j}{2m}] \delta(c_i, c_j)$$

Where $A_{i,j}$ represents the weight of the link between i and j, k_i is the sum of the weights of the links attached to vertex i, c_i is the community to which vertex i is assigned. The δ function $\delta(u,v)$ is 1 if $u=v$ and 0 otherwise, and m is the sum of weights of all the links in the network.

4 INTERNET ORGANIZATION DISCOVERY

In this section, we first show the clustering results

with real data from two RIRs, and then validate the reasonability and effectiveness of the obtained results.

4.1 Datasets and Parameter Settings

The datasets used in this paper are collected from RIPE and APNIC, both of which provide allocation registry service and routing registry service with a shared database.

Table 1: Datasets.

Data	APNIC	RIPE
22nd, Jun 2008	*AP-1*	*RP-1*
21st, Dec 2010	*AP-2*	*RP-2*
25th, Feb 2014	/	*RP-3*
10th, Aug 2015	*AP-3*	*RP-4*

Our datasets include 7 snapshots of RIR data at 4 distinct time points, as shown in Table 1. For validation purpose, we also collected 4 BGP (Border Gateway Protocol) route tables at each time points from RIPE-RIS project (collector *rrc03* is chosen for it has more peering ASes). While time diversity enables us to observe the evolution of RIR data over time, these datasets are not consistent with each other for the change of RIR policy in data release. In particular, the dataset here is largely different from that used in (Cai,2010) in the following three aspects.

- **No *org* Attributes**, including ***AP-1***, ***AP-2*** and *AP-3*. While the *org* attribute provided a good coverage on the AS objects (90% in percentage) in (Cai and Heidemann, 2010), APNIC did not use this attribute at all.
- **Partially Anonymous**, including ***RP-1***. PoC objects are anonymized by replacing the associated phone numbers, emails with +31205354444, the number of RIPE NCC and unread@ripe.net, respectively. However, the *admin-c* and *tech-c* attributes of other objects are still available.
- **Completely Anonymous**, including ***RP-2***, ***RP-3*** and ***RP-3***. Compared with *RP-1*, not only telephone numbers and email addresses, but also *admin-c* and *tech-c* attributes are removed.

As for the parameters used in the calculation of tie strength between network operators, we set a to be 1/2, and each element $s_{i,j}$ ($1 \le i, j \le n$) of matrix S is set to be 1 for simplicity. That is, no matter a link goes across two dimensions or not, it contributes the same amount of strength to the tie between network operators.

4.2 Discovery Results

The *IORG* discovery results are shown in Table 2. The first column denotes the number of vertexes in the built MDN, while the second column denotes the number of links (links generated via correlation procedure are also included).

As for the datasets, we can observe a clear trend that both the vertexes and links grow fast as time goes on, while RIPE has a higher speed. However, the number of *IORG*s in both APNIC and RIPE grow much slower than vertexes and links.

Table 2: Discovery results of *IORG*s.

Dataset	#of vertexes	#of links	#of *IORG*s
AP-1	838,143	7,459,473	56,236
AP-2	973,218	9,138,517	57,861
AP-3	1,051,342	10,496,599	57,964
RP-1	1,004,208	9,663,494	73,247
RP-2	3,824,610	38,219,328	75,336
RP-3	4,936,486	53,807,697	75,912
RP-4	5,534,401	70,021,241	76,238

Figure 5 depicts the Cdf (Cumulative distribution function) of the size of *IORG*s (the number of RPSL *person* objects in each *IORG*) from AP-2 and RP-2 dataset respectively.

As we can see, 20.0% of the *IORG*s in RIPE and 28.7% in APNIC contain only 1 *person* object, and *IORG*s consisting of fewer than 10 *person* objects account for 69.5% in RIPE and 81.3% in APNIC. That is, most of the *IORG*s are very small in scale. In fact, although there are more and more *IORG*s in the datasets as time goes, as shown in Table 2, the fraction of small size *IORG*s is becoming larger and larger. We attribute this increase to the fact that more and more users are required to register their allocations in RIR databases.

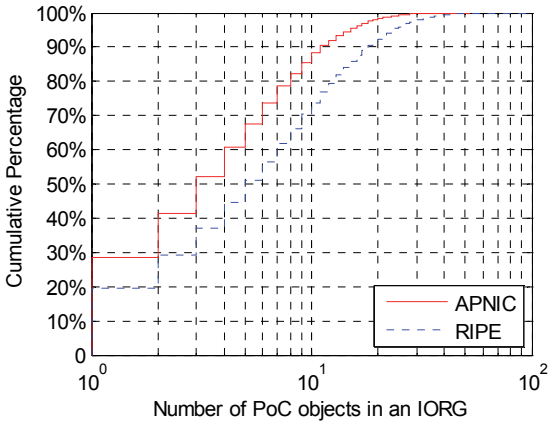

Figure 5: Cdf of the size of *IORG*s.

4.3 Validation with IP-to-AS Mappings

The IP-to-AS mappings (*prefix*, originator AS) in BGP routing reflect the up-to-date usage of Internet number resources. In an ideal case, an organization would advertise its IP prefixes with its own AS numbers, producing mappings whose *prefix* and *AS number* belong to the same organization. However, since not all the organizations are qualified to apply for AS numbers, an organization can also delegate its prefixes to ISP providers for advertisement in BGP. In this case, this organization and its providers are supposed to be relevant and close to each other in G_0.

Two indicators are defined to quantify the relativeness and closeness of the two organizations (O_{IP} and O_{AS}) that own the IP prefix and AS number respectively involved in a (*prefix*, originator AS) mapping observed from BGP route tables.

Relativeness(γ): Assuming that most of current usages are reasonable, a mapping is considered *connected* if the corresponding O_{IP} and O_{AS} are connected in G_o. Relativeness γ is defined to be the fraction of connected mappings, compared with the mappings whose O_{IP} and O_{AS} can be pinpointed in the organization graph.

Closeness(β): This indicator is defined to be the fraction of mappings whose O_{IP} and O_{AS} belong to the same organization, compared with mappings that are connected in G_o.

The validation results are shown in Table 3. Our conclusions are two-fold. First, the relativeness indicator γ is really high. That is, for most of the IP-to-AS mappings, the two organizations O_{IP} and O_{AS} have a close tie in between. This finding means that our discovery results can be used to detect prefix hijacking, routing leak or similar events, as Siganos *et al*. had done in (Siganos and Faloutsos, 2007). Second, the closeness indicator β is actually the ratio that O_{IP} and O_{AS} belong to the same organization. This is the upper bound of accuracy that traditional methods such as (Siganos and Faloutsos, 2007) can reach if they do not perform clustering operation on network operators.

Table 3: Validation results.

Dataset	γ	β
AP-1	94.5%	78.3%
AP-2	95.2%	80.5%
AP-3	94.6%	76.9%
RP-1	98.8%	84.3%
RP-2	96.5%	85.7%
RP-3	96.1%	84.4%
RP-4	97.0%	82.1%

5 CONCLUSIONS

In this paper, we develop a systematic approach to discover organizations in the Internet. Our basic idea is to discover network operators that have close ties among each other, and consider them as being from the same organization. To be honest, the model is still very coarse. However, the preliminary discovery results can enable us to start looking into the organization level Internet ecosystem.

In our future work, we would continue adjusting our model and related approach, and then extend this approach to ARIN, AFRNIC and LACNIC to obtain an organization level picture of the global Internet.

REFERENCES

Siganos, G., Faloutsos, M., 2007. Neighborhood watch for Internet Routing: Can we improve the robustness of Internet Routing today?. In *INFOCOM'07*, *26th IEEE Annual Conference on Computer Communications*. IEEE Press.

Cai, X., Heidemann, J., Krishnamurthy, B., Willinger, W., 2010. Towards an AS-to-organization map. In *ACM IMC'10*, *10th Annual Conference on Internet Measurement*. ACM Press.

Blondel, V. D., Guillaume, J.L., Lambiotte, R., Lefebvre, E., 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*.

SWIP, Shared Whois Project. http://www.arin.net/reference/database.html.

RFC2622, 1999. Routing Policy Specification Language (RPSL). IETF rfc.

RIPE-RIS Project. http://ris.ripe.net.