UDC 004.75, 004.724.2

G. V. Poryev
National Technical University of Ukraine «Kyiv Polytechnic Institute»
Peremohy ave., 37, 02056 Kiev, Ukraine

Using the Internet registries to construct a structural model for locality estimation in the overlay networks

The structure of the database files of the regional Internet registries is analyzed. With the help of proposed method for the locality class estimation is calculated the topological affinity for arbitrary address pairs without using any service traffic. The proposed locality class metric can be used as a distance substitute instead of the traditional trace and ping metrics to construct optimal overlay structure of the distributed and peer-to-peer networks.

Key words: *Internet, distributed networks, peer-to-peer networks, locality, regional registries.*

Overview

Over the course of the last two decades the number of Internet nodes has increased by several orders of magnitude. During the initial commercialization of the Internet the developments of its regional segments were mostly spontaneous. Providers were establishing links thinking of technical availability, investment amounts, marketing strategy, peering cost and many other factors none of which was related to the desired optimal structure of the regional Internet segments. As a result, the standard trace route from one ISP to another might have crossed several international borders if not continents.

The modern practice of designing the national segments in the technologically developed countries consists, in particular, in facilitating so called Internet Exchange Points (IXP, commonly abbreviated as IX) [1]. To build an IX, several mainstream national ISP will dedicate a single, territorially localized network center with routers and auxiliary equipment. The network of each ISP will then connect to such a point. Other local ISPs may then also participate in the system.

The core specific of an IX which allows reducing the maintenance cost of the links is the aggregation of member networks in such a way that the traffic from one member network to another is solely transferred through the IX. This, in turn, allows establishing mutually beneficial agreements between member providers, whereas the cost of the traf-

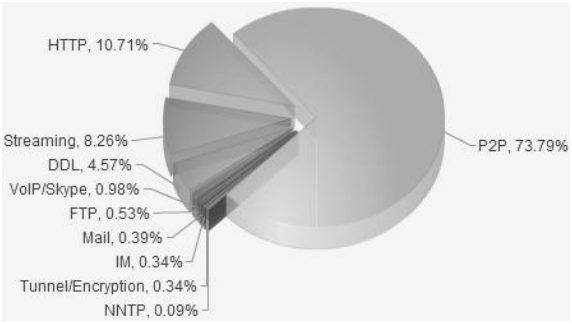
fic from provider *A* to provider *B* is assumed to be equal to the cost of the traffic from provider *B* to the provider *A*. As a result, the mutual traffic charges are effectively cancelled and the providers only pay for the bandwidth allocated regardless of the amounts of traffic consumed.

This scheme allows the ISPs to reduce the end-user prices for the traffic between its customers and IX member networks. Traditional approach is to charge no customers for the traffic of so called «internal» networks, namely member ISP networks and to apply no bandwidth shaping upon it.

The difference in guaranteed bandwidth depending on whether or not the destination node belongs to the same IX can easily reach two orders of magnitude. Relying on the cost difference, the end-users could also use private intelligent filters to minimize the «external» traffic thereby reducing their expenses should they depend on the amount of traffic consumed.

From the topological standpoint the IXs are assets with relatively high (in comparison to the average) number of member autonomous systems. For instance [1], in Ukrainian national Internet segment as of now exists two main IXs operated by the Datagroup and the Ukrainian Internet Association. Two core ASes which makes up the aggregations, namely AS21219 and AS15645, had wielded more than 35 and 50 member networks respectively as of 2008, but have now grown to aggregate more than 100 members networks each according to their web-sites. Whereas the average number of member networks in regular assets is around 3.

According to the protocol distribution study done by the Ipoque.de[1] in 2007, over the last decade the spread of peer-to-peer networks, its summary traffic almost everywhere is greater than that of all other traffic classes combined (see Figure).



Traffic class distribution for one of German backbone network

Most modern publications related to the peer-to-peer networks show little concern for the optimal overlay structures in terms of throughput and latency. For this reason, the problem of increasing the efficiency of the distributed networks on the existing transport infrastructure is of great importance.

Modern approaches to this problem, especially those using traditional locality metrics, do have some drawbacks. Traditional locality metrics such as ping response time and trace route count are influenced by the link load and conditions; the dedicated over-

[1] http://www.ipoque.de/userfiles/file/ipoque-internet-study-2007-abstract-en.pdf

lay control and maintenance infrastructures are subject to administrative shortcomings, etc.

In this work we consider building the structural model of the national Internet segment based on the information taken from open and standard sources. This model will then be used to suggest the new locality metrics without the aforementioned drawbacks.

Locality-based optimization of the peer-to-peer network

Vast majority of the distributed network users in general and peer-to-peer networks in particular are concentrated behind the «last mile», so it is reasonable to assume that the differentiation of their links by the bandwidth speeds is not as much crucial as it is for backbone networks. Average access speeds in the developed urban areas range from 1 to 10 MBits/sec of guaranteed bandwidth regardless of the destination.

In this case the difference in quality of service set up by the provider for the segments covered by the national IX whose member it is and the «outside» network is essentially a ground for optimizing the peer-to-peer overlay networks. Such systems were proposed before, but they did not draw attention enough on the part of business and are now being developed on the volunteer basis.

In network-based applications we often need global knowledge of all network nodes and distances between these nodes. This information is usually managed by a central instance or may be inferred from external infrastructures. By having this knowledge, nodes in a network are able to construct complex infrastructures and achieve efficient communication at the application as well as at the network layer.

Without relying on a central instance or external infrastructures clients usually apply either ping or hop count methods to estimate node distances. The problem hereby is, that even if the ping or hop count methods would provide reasonable and reliable results, there is no way to apply these methods to a pair of foreign IP addresses. That is, it is easy to measure the ping or hop distance from the node *a* to the node *b* or to the node *c*. But there is no way to measure the ping distance between the nodes *b* and *c* from the node *a*. That means, if clients are interested in this kind of information, they have to request explicitly this information from corresponding nodes, which causes a significant communication overhead.

To sum up, the construction of complex structures requires either additional communication between nodes (in decentralized P2P systems), or is not scalable due to the existence of a Single Point of Failure (in P2P systems with a central instance) managing relevant information, which is a big drawback in global scale networks. In order to sidestep these drawbacks, in our previous works [2, 3] we have employed decentralized P2P systems and have proposed the CARMA (Combined Affinity Reconnaissance Metric Architecture) model and metric which is calculated locally on each node.

This metric is calculated given the remote IP address of the peer and all information than can be implicitly inferred from it. The «combined» adjective in the CARMA acronym means that despite our work's prevailing focus on only the first layer of the proposed metric, its design does, however, contain several additional components that can be included in a metric calculation in the future, as follows: 1) average response time to keep-alive requests; 2) average hop count to the destination, including the possibility of its change during communication; 3) bandwidth and average consumption at

the moment of decision, including preset constraints; 4) «gratitude» and «greed» values calculated as the amount of traffic the remote party had provided and consumed respectively; 5) any other parameters relevant to the node pairing or the overlay network as a whole.

Generally speaking, we consider CARMA as three layered, with the first layer being the locality awareness expressed in classes, a second layer that utilizes additional traffic but does not involve actual P2P communication, and a third layer that requires active communication to remote parties over a compatible protocol.

Of course, CARMA metric can be calculated for an arbitrary pair of allocated IPv4 addresses, which is one of its advantages over competing methods. But in practical scenarios utilizing locality class to build optimized data exchange one of the addresses will almost always be that of local host, i.e. host running the CARMA instance itself. Whereas the ability to calculate the locality class for non-local pair of addresses may, for instance, be used to facilitate observation and quality assessment service within the overlay network.

CARMA works by initially preloading structural information from publicly accessible services called Regional Internet Registries (RIRs) and converting it into an internal graph-like data structure. Unlike solutions based on the PlanetLab infrastructure or those using RouteViews, the RIR services and databases are mandatorily public, essential for the functioning of the Internet and therefore much more reliable. The pieces of information important to CARMA include the delegated-latest-* databases of registered IPv4 ranges, Autonomous System Numbers (ASNs) and various WHOIS databases of registered sub-ranges and Autonomous System Sets (ASSETs).

It should be noted that although the information stored in RIR databases may resemble to some degree topological junctions available from BGP, its purpose was never to maintain the real-time track of the actual precise Internet topology. The CARMA model is designed to estimate (not measure or calculate) the relative locality class, therefore it permits some tolerance towards latency or inconsistencies of its sources, namely RIRs.

The RIR information reflects the topological structure of the Internet as it was designed by the maintainers of its segments at the moment they submit relevant information to RIR. Assuming good faith and high skills of the personnel responsible for network maintenance we should expect the actual topology to be very close to what is in the RIR database.

Knowing very well that the topological structure of Internet is not static by nature and that the overloading, balancing, hardware and software failures sometimes cause the information flow switchovers, we however understand that the occurrences of this problems are generally low. Therefore they cannot affect the topological structure on a scale wide enough to cause CARMA model to begin returning massively erroneous results on a large (on the order of hundreds and thousands) peer lists, whereby lower percentages of incorrectly estimated localities are rather acceptable in most peer-to-peer scenarios.

To sum up, the fact that daily RIR database information may not reflect the immediate state of Internet topology does not affect CARMA performance in general.

The RIR database expansion rate is usually relatively slow — about tens of IPv4 ranges allocated daily. Since IANA's allocation of the last IPv4 range to the RIRs at the

beginning of 2011, the expansion rate is even slower and IPv4 allocations will soon stop completely. However, before the IPv6 gains worldwide spread, IPv4 allocations are expected to make up the majority of Internet addresses in some years.

In spite all of this, CARMA only needs to update its locally cached data from RIR databases once in several days. Once loaded, CARMA builds a model to approximate the Internet topology with some simplifications, resulting in 4 structural layers as follows: a) IPv4 ranges are divided into b) sub-ranges but at the same time they also belong to c) Autonomous Systems (ASes), which are joined into sets called d) AS-SETs or ASSETs.

It is assumed that lower layer entities are explicitly connected through their common upper layer entity, and ASSETs are arbitrarily connected to each other. It is understood that such assumptions in the model are more optimistic than what happens in reality, as there, for example, may be ASes that spread worldwide. However, exceptions like this are not numerous and pertain only to the departments of the few ISPs specializing in providing transoceanic and transcontinental backbone links. Therefore, Internet end-users are unlikely to be encountered in the ranges assigned to such ASes.

Analysis of the RIR database

The routers comprising the backbone, data-centers, big corporate networks with significant presence in the Internet use the Border Gateway Protocol (BGP) protocol to arrange rules according to which incoming packets are processed and redirected to corresponding network segment. The BGP protocol includes, in particular, the announcing mechanism, whereby the trusted router distributes the message that it is ready to serve a particular IPv4 address space segment and such a message then gradually reaches other routers. They then add an announcement to their internal routing table according to the uplink capacities.

Having access to the BGP information which reflects the present link status it is possible to estimate distance metric with high precision, however the BGP is only used for the multi-homed routers in the backbone networks and in the geographically distributed enterprise networks.

Hence for the task of estimating the topological locality we can use the database files published by the RIRs, which contain all the necessary data concerning the IP address ranges, autonomous systems definitions and their relation.

For case study we use the IPv4 address 77.47.192.74, which belongs to the address spaces in such an order: a) general IPv4 address space (2^{32} nodes); b) European address space managed by RIPE RIR (about 50 % of the entire address space); c) specific solid IPv4 range allocated to the organization entity registered in Ukraine, according to the *delegated-* database file; d) address block allocated for NTUU «KPI» needs; e) eight IPv4 addresses allocated for one of the departments of the university.

Such a scenario of the sequential allocation is typical for end-users in the educational and corporate sector, including hosting and telecommunication services. Home and office sector only differ by having end-user IP pool configured to dynamically allocate address to most client connections.

Using the aforementioned address 77.47.192.74 belonging to the web-server of the department of NTUU «KPI» as reference point, we shall research how, using the RIR database files, the information relevant to locality can be inferred.

Let us discuss the structural layers based on the RIPE RIR information in more detail.

IPv4 range is a subset of an IPv4 address space defined by the first address of the range and a host count. Note that the host count is not necessarily a power of 2 as implied by the Classless Inter-Domain Routing (CIDR) rules now commonly used for the Internet routing. There are records that specify an arbitrary number of nodes, but for practical reasons such definitions are subsequently augmented by CARMA to contain a 2^n, $n \in N$ number of nodes. IPv4 ranges are defined in the *delegated-* file, where our reference address belongs to the range originated at 77.47.128.0 with $2^{15} = 32768$ hosts in it assigned to Ukraine (ISO country code «UA») on January 15, 2007.

AS is a registered Autonomous System. Every AS has a numerical identifier known as Autonomous System Number ASN. AS definitions are also listed in the *delegated-* file along with the ISO country code and the date of allocation. However, this file does not specify a relation between IPv4 ranges and ASes. For this relation CARMA uses the *ripe.db.route.gz* file.

The latter file contains definition blocks, each block specifies an IPv4 range (this time in proper CIDR notation), and related ASes. This information is used to establish relations between ranges and ASes listed in the *delegated-* files. Note that a relation between an IPv4 range and an AS is not unambiguous: the same range can be announced under different ASes; some ASes or ranges listed in the *delegated-* file may not be linked at all, and some relations specified in *ripe.db.route.gz* may contain ASes and IPv4 ranges, which are unspecified in the *delegated-* file. The incidence of such inconsistencies is low, however. From *ripe.db.route.gz* file one can see that NTUU "KPI" was registered in RIPE at December 11, 2002 as an AS25500 within Ukrainian Internet segment. To this AS the IPv4 range is attached, specified as 77.47.128.0/17, whereby «/17» in CIDR notation denotes the number of allocated addresses. The relation between 32768 and «/17» is established as follows: $2^{(32-17)} = 2^{15} = 32768$, since the number of bits in the IPv4 address is 32.

IPv4 sub-range is a subset of the IPv4 address space defined by the addresses of the first and last address of the sub-range. These definitions are listed in the *ripe.db.inetnum.gz* file. The sub-ranges differ from ranges in that they are not explicitly related to an AS. Sub-ranges are generally smaller in terms of address space. A vast majority of them are derived from splitting up ranges. It is therefore possible to establish a relation between one or more sub-ranges and a single range, although not all ranges are split into sub-ranges. When parsing information from this file, one should take care to check for sanity of the sub-ranges specified. For instance, a sub-range may specify an entire IPv4 address space, or a sub-range may even have a netmask length such as 3 bits and may thus be much larger than an IPv4 range. Such cases are dictated by the internal works of the WHOIS server software but are obviously invalid for CARMA and therefore filtered out of the model.

The definition concerning our reference address is listed in *ripe.db.inetnum.gz* file using the range notation (77.47.192.72 – 77.47.192.79), including an additional administrative information and date of allocation: November 3, 2008.

AS set or ASSET is a topological junction point that may declare an arbitrary number of ASes and other ASSETs and facilitate connectivity among them. It is assumed that the information flow between two ASes belonging to the same ASSET does not take a route via other ASSETs. Unlike ASes, ASSETs have alphanumeric identifiers. In terms of CARMA, an IX point is an ASSET with a significant number of member ASes (usually hundreds), although, technically, each ASSET can be considered as a kind of IX as there is usually no explicit requirement in terms of member count. The definitions for ASSETs can be found in the *ripe.db.asset* file.

In this file one can see many instances of AS25500 being declared as a member of various assets, including AS-UAIX managed by UkrTelecom and other notable backbone ISPs such as URAN network and ColoCall ISP.

It should be noted that the definitions in the RIPE RIR database are administrative rather than purely technical, and their public availability is the consequence of the openness and transparency policy of the Internet governing structures.

We have proposed to use the RIR database as the source of information to build a topological structural model.

During the design stage for the applied locality estimation system we have discovered that there is no uniform standard regulating the format of RIR databases, and each of 5 world RIRs uses its own format depending on the version and vendor of the WHOIS-server.

It was also found out that some RIR sites do not contain the entire locality relevant information. For instance, AS definitions are available only at RIPE and AFRINIC RIRs. At the same time, IP range definitions are available from all five RIRs.

However, for the proposed method of locality estimation all those missing pieces of information can only mean fewer locality classes for the affected segments. In the worst case, should there be only *delegated-* file, the method can still return two classes — «range» and «horizon». Fortunately, the RIRs serving the regions with the most developed Internet infrastructure and markets, namely RIPE, ARIN, APNIC provide more detailed information, thereby we consider the proposed method as viable and efficient in the most application scenarios.

When all database files are processed, the resulting incomplete graph reflects the Internet segment topology as close as it could be done without having access to BGP information. It is not necessary to devise any graph-walking algorithm for calculating the locality class, because the purpose of CARMA is to estimate the affinity of two given nodes, not calculating an exact hop count.

Collisions and inconsistencies in RIR databases

It should be noted that RIR databases define an IPv4 ranges from two different sources and these databases are, as noted above, administrative, not technical. Therefore they may contain certain errors and omissions contributed by the human factor and policy considerations.

Let us discuss in more detail the *delegated-* file, which is mandatorily served by all five RIRs and, unlike other databases, has the same standard format. In its definitions we found no IPv4 range intersections but there are inconsistencies with geopolitical assignments.

For instance, the IPv4 block «193.111.2.0/23» in the *delegated-* database is defined as belonging to the business entity registered in Ukraine, whereas the database file *inetnum* (which is correct in this case) points out that this block belongs to Microelectronics, Ltd, Russia.

Should some method of determining the national flavors of IPv4 addresses rely on data derived from *delegated-* file, it would return an erroneous geopolitical assignment of an IPv4 range.

However, just the same internal inconsistency of RIR database cannot affect the building of the topological structural model of the Internet using our proposed method. It is achieved since at the stage where AS and ASSETS linkages are inferred it was found out that the specific range does not belong to the Ukrainian IX, no matter if the range was indeed owned by an Ukrainian business entity.

The reverse situation is also possible, dictated by the administrative policy of the ISP. For instance, some IPv4 ranges allocated to end-user IP pool from UkrTelecom (such as 95.132.0.0/16) are thought to be operating within the scope of an UA-IX exchange point, based on the information inferred from RIPE RIR. However, the explicit trace-route check from this range towards almost any host «covered» by UA-IX or Datagroup IX indicates the explicit routing via the foreign ISPs, especially Level3.

Moreover, the *inetnum* database contains some sub-range definitions which are syntactically valid but invalid for inferring the topological structure of the Internet.

For example, the *inetnum* database contains the definition of an entire IPv4 address space in the form of «0.0.0.0/0» record. This is dictated by the specifications of the WHOIS server, which is required to return a reply even if the IPv4 address requested does not belong to its «responsibility».

There are also definitions which are subsets of each other, such as delegated block 77.47.128.0/17 in *delegated-, route and inetnum* files. In the *inetnum*, however, there is also a sub-range 77.47.128.0/22 which is a subset of the latter.

Since the *inetnum* database structure does not guarantee that the subsets are defined always after (or always before) their respective parent sets, the search for the sub-range binding has to be conducted bearing in mind that the first encountered definition may have not enough bits in its network mask to get necessary precision in estimating the locality class.

Generally speaking, the mentioned and other typical inconsistencies of the RIRs can be classified such as:

— those not affecting the correctness of the locality estimation (excessive information, over-spanned ranges);

— those affecting the locality estimation, like omitted ranges, null linkage etc.

However, it is understood that the RIR databases are maintained in their integrity, as well as the structural models of the Internet segments based on it. The method of locality estimation always returns a result. Moreover, the error caused by database inconsistencies and omissions affecting lower classes of locality can only lead to locality be-

ing returned as one class up, and not to the crash of the whole algorithm. But even if the severe errors in the RIR database lead to the crash and erroneous return of «horizon» class for the address pair, the nature of peer list in peer-to-peer networks is such that there is always a node of close locality class for which the estimation algorithm would return correct class.

Additionally, the practical implementation of the locality estimation method is designed to include a second layer of traditional locality metric estimation techniques, such as ping and trace-route; hence the probability of both layers to fail simultaneously on same IPv4 address is negligible.

Our estimation suggests that even failure of all three layers of locality estimation for less than 10 % of peer list would not affect the efficiency of data transfer with the overlay network. Modern peer-to-peer clients for which this method is primarily designed are operating on the peer lists consisting hundreds and thousands of addresses, therefore the majority of nodes unaffected by the inconsistencies is still enough to maintain full load on the dedicated bandwidth.

Conclusions

1. The proposed method of deriving the structural model of the national Internet segments and the method of estimating the locality classes based on the latter are, in general, tolerant to collisions and inconsistencies in the RIR databases causing the algorithm to return incorrect locality class for the IPv4 address pair.

2. Implementation of the proposed topological locality estimation system with the pre-computed structural model of the national Internet segments will allow for automatic adaptive clustering of the overlay network using the peer list reordering based on the locality class. Unlike the traditional metric, proposed method at its first implemented layer does not use service traffic to measure directly topological distance and therefore can be used to estimate the locality even for non-local address pairs.

1. *Фурашев В.* Параметры украинского сегмента Internet как сложной сети / В. Фурашев, В. Зубок, Д. Ланде // Открытые информационные и компьютерные технологии: сб. науч. тр. — 2008. — Т. 40. — С. 235–242.

2. *Poryev G.* Multi-Tier Locality Awareness in Distributed Networks / G. Poryev // Інформаційні технології та комп'ютерна інженерія. — 2009. — № 3(16). — С. 13–17.

3. *Poryev G.* CARMA Based MST Approximation for Multicast Provision in P2P Networks / G. Poryev, H. Schloss, R. Oechsle // In Proceedings of the Sixth International Conf. on Networking and Services (ICNS 2010) / IEEE. — Cancun (Mexico). — 2010. — P. 123–128.

Received 22.02.2011