УДК 004.75, 004.724.2

G.V. Poryev
National Technical University of Ukraine "Kyiv Polytechnic Institute", Kiev
E-mail: core@barvinok.net

NEW METHOD OF RELATIVE LOCALITY ESTIMATION ON THE INTERNET

Abstract
Poryev G.V. New Method of Relative Locality Estimation on the Internet. *In this paper we have shown that the existing methods of estimating relative locality of network nodes are far from providing diverse and accurate results. The proposed improved method is based on the information inferred from regional Internet registries and is a part of the more complex distance substitute metric.*
Keywords: *peer-to-peer networks, distributed networks, locality, distance metric.*

Introduction. Today, a diverse variety of peer-to-peer (P2P) networks exist. Among the best known are eDonkey2000 (ED2K), GNUTella and BitTorrent [1]. Regardless of the differences in their protocols and implementations, there are common procedures in all file-sharing networks. After the request for published entity is processed by either the indexing server or other nodes, a response comes, containing a list of bootstrap peers — network nodes that may serve the requested content. Whether it is done using DHT such as Kademlia [2], using indexing server or message flooding, the result will contain at least the list of IP addresses and ports. From this point on, it is completely up to client software to decide which nodes should be queried and in what order.

The analysis of ED2K and BitTorrent network traffic from a single node indicate that client software usually performs queries in an unsorted order initially reported by network or index server. In the popular BitTorrent tracking servers, the number of peers for highly demanded content could easily reach tens of thousands, whereas for most end-user nodes it is quite impractical to initiate more than hundred connections simultaneously.

The core of the proposed idea is to not leave the peer selection process to a pure luck, but rather assign a calculated priority to each peer by introducing special relative locality metric in the querying order, which, we believe, may significantly increase overall performance of the network.

Analysis of modern solutions. ED2K clients will query every known source and will attempt to place themselves in the download queue of every source they managed to successfully negotiate with. The other (receiving) side will organize the download queue initially according to first-in-first-out principle. Modern clients, such as eMule, also feature reward system, which advances inbound client in the queue according to the amount of related traffic they had provided to the node. This is supposed to discourage leeching but also have obvious drawback in delaying new nodes that do not have any part of the content yet. Although eMule provides a few tuning methods such as queue rotation speed and chunk management based on file's popularity, none of it takes into account anything related to connectivity (client bandwidth, network latency etc). The big difference between BitTorrent protocol and eDonkey2000 in this regard is that it does not feature any reward system, and due to per content swarm isolation BitTorrent is generally faster. Also, tracker may initially not report all peers to the client. However, this may be circumvented later by the peer exchange and DHT mechanisms. Recently there have been some advances in the relative locality awareness for BitTorrent networks. Popular nationwide trackers have introduced so-called "retrackers" — dedicated secondary servers optionally connected to primary database but mainly supposed to only return peer list local to specific network scope — usually within IP address pool allo-

cated to customers of a particular ISP. This provides for significant speed burst for affected ISP clients, but it is very simple method that only allows for two-tier relative locality awareness.

Topological study of the Internet. Depending on the scale of observation, Internet may be considered as either decentralized network or a cluster of centralized network segments. Basically, Internet contains large backbone networks involved in international and intercontinental links, national-tier ISPs, end-user-servicing ISPs, hosting companies and, of course, end-users. Network latency and quality of service are accordingly very different depending on the link speed from tens of Gbps to the speeds of dialup modems, less than 56 Kbps. On the scale of a country, Internet structure used to be organized rather sporadically — individual ISPs established arbitrary links among themselves and to foreign upstream ISPs. This had lead to peering conflicts and situations in which a message to a neighboring house traveled halfway the continent. To alleviate this problem Internet Exchange points (IX) were introduced. Usually, a number of national telecom operators create the dedicated facility to which all national ISPs then connect. Thus consumer traffic within the scope of IX does not travel expensive international or satellite links, and this helps balance mutual peering, ensure lower costs of maintenance per ISP therefore allowing lower costs to end-users. Some developed countries have more than one national IXes.

The network segment covered by an IX does not form strict hierarchical graph, but it is nonetheless complete, even if more than one AS must be traversed to reach IX routers. These situations, however, are rare, except for stub ASes (AS linked to only one upstream AS). From the customer point of view it is generally assumed that traffic within single IX flows faster and is cheaper than external. Also, IX can provide for lower number of hops a packet must traverse. Modern network modeling environments that deal with Internet topology rarely take relative locality into account. Most of them use either network latency metric measured in time units between request and response (also known as ping-pong) or hop count metric measured as number of nodes between source and destination hosts [3]. Not only these measured parameters are subject to unpredictable changes due to network switchovers, but also they involve additional traffic, considerable sampling time and may sometimes not yield any results at all. This is usually caused by some kind of traffic filtering — for example, some core routers of large enterprise network filter out all incoming ICMP packets. Additionally, we deem ping method as generally unreliable as it heavily depends on link speeds and bandwidth conditions, for example, zero-loaded end-user ADSL link can produce slower pings than almost fully loaded Gbps link. Also, as shown in [4] standard trace-route methods may also be unreliable and affected by bandwidth conditions or indicating nonexistent links due to traffic switchovers, which are, in turn, may be caused by widely implemented load-balancing techniques that encourage packets to travel multiple routes. As [4] goes, «where there is load balancing, there is no longer a single route from a source to a destination. In the case of per-packet load balancing, a given packet might take any one of a number of possible routes. With per-flow load balancing, the notion of a single route persists for packets belonging to a given flow, but different flows for the same source/destination pair can follow different routes. Designing a new trace-route able to uncover all routes from a source to a given destination would be a significant improvement. Classic trace-route is not adequate to the task, as it cannot definitively identify one single route from among many. It suffers from two systematic problems: it fails to discover true nodes and links, and it may report false links.

These problems arise because trace-route discovers hops along a route with a series of probe packets, and the fact that a load-balancing router, or load balancer, can direct these probes along different paths», see Fig.1. With Paris trace-route [4], Augustin et al. have proposed a solution that deals solely with the side-effects caused by load-balancing by improving the trace-route specifically to circumvent packet similarity. It is known that load-balancing routers tag incoming packets with so called “flow identifier” according to the packet header content, including checksum. Whenever any packet header fields vary, the packet might be assigned different flow identifier and therefore

undergo different routing, resulting in different traverse path with all aforementioned anomalies caused by this.

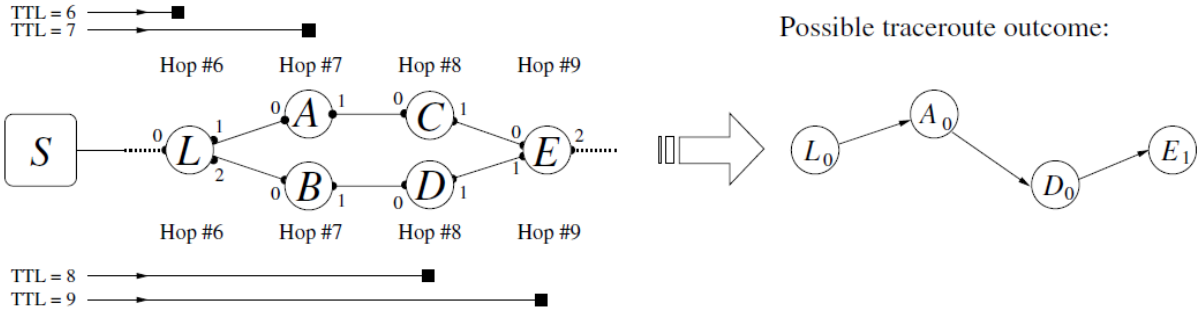

Fig.1. False network topology inferred from ICMP responses (per [4]).

The Paris trace-route method takes special steps to ensure that packet header contents remain constant throughout the whole probing. Historically, trace-route tools had used three most widespread protocols for path probing, namely TCP, UDP and ICMP.

While the latter might seem to be most fit for the purpose as having been naturally designed for network management, it is sometimes filtered on borderline routers, rendering such trace-route method ineffective. This is why the default protocol used for trace-route probing in GNU environments uses UDP.

If no special steps are taken, each consecutive packet of trace-route differs in packet header content. For ICMP and TCP, the Sequence Number varies; for UDP the content varies causing checksum to change for each packet. Paris trace-route thus manipulates packet content and sequence number fields to ensure each subsequent packet header is the same as previous. And by doing this, it assures the probe will undergo the same routing, and unlikely be subject to traffic switchovers. Notwithstanding the innovative solution that Paris trace-route presents, experimental results indicated that although more accuracy was indeed achieved in measuring traffic topology, Paris trace-route is still subject to artifacts such as trace loops, cycles and diamonds with surprisingly higher percentage of occurrence.

It should be also pointed out that like standard trace-route, Paris trace-route require at least the same amount of time and traffic to yield topological information. While in some cases this is affordable, our peer-to-peer optimization generally demand low latency solution for quite high number of arbitrary peer nodes. Even if each of the nodes is to be probed by trace-route simultaneously with all others, it will still take up to a minute and considerable service traffic overhead to sort out their priorities. Newly appeared peer nodes must also undergo the same estimation while the traffic exchange may have already been active for long, exhausting available bandwidth and slowing the estimation further.

New method for locality estimation. To alleviate the problem and provide more reliable method (at the cost of CPU load and significant startup time), we propose the combined relative locality metric (CRLM) which is calculated locally on each node independently of the others and is only meaningful within the scope of this node.

Metric is calculated given the remote IP address of the peer and all information than can be inferred from it, including previously stored and all information the remote peer can report on itself, if such feature is provided by the application layer protocol. Components of the metric include the following conditions: whether remote party belongs to the same subnet, same AS or same IX; average response time to pings (not necessarily standard ICMP, but rather some internal keep-alive specific to protocol); average hop count to the destination (including the possibility of change [4]); bandwidth and average consumption at the moment of decision, including preset constraints (must be reported by the remote party, if possible by protocol means); “gratitude” and “greed” values cal-

culated as the amount of traffic the remote party had provided and consumed respectively (note the ED2K reward system above).

Obviously, only the first component of the CRLM means true relative locality awareness which can be derived without prior communications to the remote party, relying purely on pre-initialized internal topology database. The nature of CRLM is three-layered, with first layer being the aforementioned true relative locality awareness, the second layer that utilizes additional traffic but does not involve actual P2P communication (second and third components) and third layer that require active communication to the remote party over compatible protocol (fourth and fifth components). CRLM is meant to be dynamically changing as the communication goes, reflecting and adapting to the changes in bandwidth conditions and to the current network load. Life-cycle of CRLM-capable node in P2P network may start with initiating startup sequence upon achieving connectivity to the Internet. The startup sequence involves downloading most recent IP and AS allocation databases from all five worldwide RIRs and compiling them into easily indexable internal format. This may take few minutes to complete, depending on the CPU speed and bandwidth. Although the expiration term for such data is usually one day, the rate at which new AS and IP ranges are being allocated is not significant and therefore startup sequence may be called to refresh data less frequently than on a daily basis. This approach of pre-downloading database excerpts relevant to our relative locality calculations is intended to achieve independent knowledge of topological information. Many researchers, such as [5] tend to consider the BGP protocol as almost the only reliable source of information on the relationship between IP and AS. And indeed it is, but only when the peer node in question have the real-time access to the most recent BGP snapshots or specially deployed web services that provide most recent summarized routing state. For laboratory research or academic purposes it is easily achievable by administrative means. But for end-user community involved in peer-to-peer information exchange, especially file sharing, it is not. We understand that the information stored in regional Internet registries pertaining IP and ASN allocations and their linkage is only an approximation. But it is a very close approximation that once downloaded and processed, allows peer-to-peer client to use no overhead traffic at all and come up with a sorting solution immediately. The result of the first layer of CRLM is a flavor of target node in relation to originator node. Proposed flavors are:

1. ”Same subrange” identifies the presence of remote node within the same IP subrange as defined in WHOIS database excerpt dealing with administrative IP subranges. This is most likely within the scope of operation of single router; for example, this could be end-users connected to the same POP (Point of Presence) of telecom operator, or nodes within the university network, which usually have single upstream ISP;
2. “Same range” identifies the presence of remote node within the same range IP range as defined in WHOIS database excerpt dealing with ASN and IP delegations. This is most likely within the scope of the same department or small organization.
3. ”Same AS” identifies the presence of remote node within the address space allocated to the AS (Autonomous System); although this does not guarantee such immediate connectivity as previous states, packets are unlikely to travel networks outside AS since AS is the basic Internet routing entity [6] (all network destinations announced by an AS undergo the same routing rules everywhere) and are handled by ISP internally;
4. ”Same ASSET/IX” states that both originator’s and target’s nodes belongs to the same AS-set, which may happen to be IX if the number of member links are significant; the immediate advantage of this knowledge is not obvious but it in the developing countries difference in quality of service may largely depend on this flavor to the point that network speed and latency differ by the orders of magnitude;
5. ”Distant” identifies that the relative locality of originator’s and target’s nodes cannot be reliably estimated and therefore they assumed to be located topologically farthest away.

Certainly, the true impact of the CMP implementation on the network efficiency cannot be reliably measured until this functionality makes its way into existing popular clients of p2p networks. However, estimations could be done, assuming that nodes with specific relative locality flavors will provide for significant speed bursts.

Experimental results. The test surveys were conducted on the swarms formed around content units published on the private Russian BitTorrent tracker Torrents.Ru. The choice of this tracker instead of its Ukrainian counterpart was largely due to the abundance of nodes the latter have under UA-IX coverage, which could render improvements minimal, if any.

Choosing any foreign tracker to conduct a test survey is also unlikely to yield any significant results because the maximum meaningful CRLM flavor ends on a national level, therefore almost all peers will be flavored “distant”. Torrents.Ru, however, historically harbors great community of Ukrainian users, which, however, is roughly less than third of total user base, according to the observed frequency in the indicated locality within the tracker’s forum. This makes that tracker perfect testing ground for our survey runs. The survey consists of accumulating 6000 IP addresses per 20 highly-active swarms and calculating CRLM flavor of local node (belonging to the user pool of ISP “Volia-Cable”) against each of them. The results are shown on Fig.2.

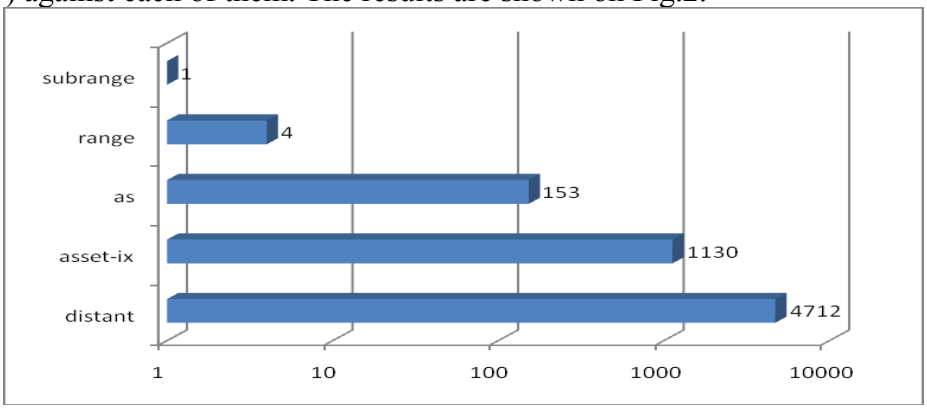

Fig.2. Flavor distributions on test survey (logarithmic scale).

The results suggest that given traffic shaping relevant to the presence within UA-IX coverage, if non-distant peers were queried prior to distant ones, the originator node might have had no need to contact distant peers at all, having exhausted its inbound bandwidth capacity with such “filtered out” high-speed peers. It is also interesting to note that the number of active peers under UA-IX coverage is significantly less than. Of course, since the results were average on 20 arbitrarily chosen swarms, there is no guarantee that the effect may manifest itself in every case.

Conclusions.

In this work we largely focused on the first layer of CRLM that consists of only one of the proposed components, for it provides benefit from preliminary topological knowledge and does not involve any measurement-related traffic. In our future work we would like to address the second and third layers of CRLM calculated by direct measurements involving additional traffic.

These layers may be expressed as weighted scores by which all peer priorities are then fine-tuned within the boundaries of their respective first-layer flavors. It should be noted, that implementation of second CRLM layer will require modifications to the existing software and third CRLM layer will require creating extensions to existing protocols in order to have any impact on the performance. In this case, the life-cycle of CRLM-capable node is extended to include the following steps after initial startup and peer list ordering in terms of first-layer flavors:

1. an additional check is performed using second layer of CRLM, in such a way as not to substitute the original order but rather to fine-tune it;
2. at this point, the actual communication to the remote party is initiated; if connections are being made using CRLM-enabled protocol, third layer is utilized by the parties providing each other

with bandwidth conditions information and related settings; using this information, peer list may once again be reordered to place less-loaded nodes closer to the subsequent queries.

If peer exchange mechanism is enabled, newly reported nodes must go through all layers of CRLM to be placed in the peer list. We plan to demonstrate the effectiveness of CRLM approach by implementing it in real-life P2P networks and extensive experiments, for which we are developing the software library implementing CRLM method under LGPL license to assist software engineers wishing to optimize performance of their P2P applications. This research was conducted in the framework of the Presidential Grant for the Support of Scientific Research of Young Scientists; project number GP/F27-0040.

References

1. B. Cohen, “Incentives build robustness in BitTorrent” // First Workshop on the Economics of Peer-to-Peer Systems, 2003.
2. Maymounkov P., Mazi`eres D. Kademlia: A Peer-to-peer Information System Based on the XOR Metric // Electronic Proceedings for the 1st International Workshop on Peer-to-Peer Systems (IPTPS'02).—MIT Faculty Club, Cambridge, MA, USA, 2002.—P.20-25.
3. G. Lucas, A. Ghose, and J. Chuang, “On characterizing affinity and its impact on network performance” in MoMeTools’03: Proceedings of the ACM SIGCOMM workshop on Models, methods and tools for reproducible network research. New York, NY, USA: ACM, 2003, pp.65–75.
4. B.Augustin, X.Cuvellier, B.Orgogozo, F.Viger, T. Friedman, M. Latapy, C.Magnien, and R. Teixeira, “Avoiding traceroute anomalies with Paris traceroute” in IMC ’06: Proceedings of the 6th ACM SIGCOMM conference on Internet measurement.New York,NY,USA: ACM,2006, pp.153–158.
5. Z. M. Mao, D. Johnson, J. Rexford, J. Wang, and R. H. Katz, “Scalable and accurate identification of AS-level forwarding paths,” in Proc. IEEE INFOCOM, March 2004.
6. J.Hawkinson and T.Bates, “Guidelines for creation,selection, and registration of an Autonomous System (AS)”,RFC 1930 (Best Current Practice), Internet Engineering Task Force (IETF), Mar.1996.

Аннотация

Порев Г.В. Новый метод оценки относительной локальности в Интернет. *В этой работе нами показано, что существующие методы оценки относительной локальности узлов сети далеки от предоставления точных и диверсифицированных результатов. Предлагаемый усовершенствованный метод основан на информации, получаемой из региональных реестров Интернет, и является частью более сложной метрики дистанции.*

Ключевые слова: *одноранговые сети, распределённые сети, локальность, метрика расстояния.*

Анотація

Порєв Г.В. Новий метод оцінки відносної локальності в Інтернет. *В цій роботі нами показано, що існуючі методи оцінки відносної локальності вузлів мережі далекі від надання точних та диверсифікованих результатів. Запропонований вдосконалений метод базується на інформації, яка надається регіональними реєстрами Інтернет та є частиною більш складної метрики дистанції.*

Ключові слова: *одноранговi мереді, розподілені мережі, локальність, метрика відстані.*

Здано в редакцію: 24.02.10р.

Рекомендовано до друку: д.т.н, проф. Зорі А.А.