Final project report to the NKFIH Grant No. 129589, “Routing in large scale networks” (Hungarian title: “Forgalomirányítás nagyméretű hálózatokban”) under the KH-18 funding scheme

Research objectives of the project

Communication is one of the most basic functions of any network. Even though different disciplines use different terms, transportation, transshipment, signaling, packet forwarding, or searching, these terms all mark the same thing: that a well-constructed signal propagation path is indispensable for a network element to interact with another one that is not directly connected to it. Without such a path available, network elements are isolated and the network is no longer a functional communications infrastructure, just a bunch of unrelated actors. In real-world networks endpoints usually employ multiple parallel signal propagation paths simultaneously to reach each other: in social networks information and rumors normally spread along multiple independent "friendship" chains, on the Internet communications typically occurs along multiple forwarding paths, etc. Currently, there is very little understanding available as to how (and why) multipath communication routes emerge in large-scale networks. Most theoretical work presumes that communication occurs explicitly along the single shortest (i.e., the least costly) path between a source and the destination, even though empirical proof exists that on the Internet, for instance, communication paths are usually longer ("inflated") than the absolute shortest possible. The ultimate goal of this project was to uncover the underlying rules of multipath routing in large networks, based on a solid understanding and model of the related path selection mechanism, and the application of the results to solve important real-world problems.

The project focused on one of the most critical issues facing today's large-scale natural (biological, social, or human-navigation) and artificial (the Internet) networks, namely the question of scalable communication schemes. Numerous studies have argued that routing and signaling schemes, and the theoretical models that describe them, which are usually restricted to a single communication path between sender node and sink nodes and are optimized for small-scale networks, will not be able to cope with ever-extending future networks having inherently multipath nature. In this context, extension stems from the natural growth of the network, and multipath communications originate from nodes' need for resilience, efficiency, signal quality, and security. The main goal of the project was to develop an empirical understanding and an analytical model for multipath routing in large-scale networks. In the project, we uncovered the essential characteristics of signal propagation in large-scale networks, we developed a mathematical model of real-world routing policies and systems, and we made predictions regarding the scalability and performance expected in growing real-life natural and artificial networks.

Research results

I'm really proud of my research team as we managed to fulfill all our plans to an appropriate extent. Our research work during this project contributed to basically three fields: (*i*), reliable multipath data transmission in the large-scale topology of the Internet, (*ii*) a theoretical model for routing scalability, and (*iii*) study and modeling of human navigation schemes.

Reliable multipath data transmission in the large-scale topology of the Internet.

The Internet can be considered as the global internetwork of independent subnetworks, or Autonomous Systems (ASes), and the goal of the Internet inter-domain routing system (BGP) is to identify and, using local business-oriented or economically motivated path selection policies, to optimize traffic forwarding paths through this large-scale AS-level topology. Our objective was to identify these multipath communication structures on the Internet, by applying a minor modification to the inter-domain routing system. Nowadays, a majority of Internet service providers are either piloting or migrating to software-defined networking (SDN) in their networks. In an SDN architecture, a central network controller has a top-down view of the network and can directly configure each of its physical switches. It opens up several fundamental unsolved challenges, such as deploying efficient multipath routing that can provide disjoint end-to-end paths, each one satisfying specific operational goals (e.g., shortest possible), without overwhelming the data plane with a prohibitive amount of forwarding state. We studied the problem of finding a pair of shortest (node- or edge-) disjoint paths that can be represented by only two forwarding table entries per destination. Building on prior work on minimum length redundant trees, we showed that the complexity of the underlying mathematical problem is NP-complete and we presented fast heuristic algorithms. By extensive simulations, we showed that it is possible to very closely attain the optimal path length with our algorithms (the gap is just 1%-5%), eventually opening the door for wide-scale multipath routing deployments. Finally, we showed that even if a primary tree is already given it remains NP-complete to find a minimum length secondary tree concerning this primary tree. The results are published as a JSAC article:
Tapolcai János, Rétvári Gábor, Babarczi Péter, Berczi-Kovács Erika: Scalable and Efficient Multipath Routing via Redundant Trees, IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, 2019 IF: 9.302

Turning the Internet into a multi-path environment could solve many difficulties network operators are facing today. There are already solutions to configure an IP network to offer multiple partially disjoint paths towards the destination. In a demo study, we focused on the performance of how multipath TCP can distribute the traffic along these paths. When the routes are not fully disjoint their throughput could be limited by some bottleneck links. Because of this dependency finding the maximal throughput in multipath TCP may call for solving a complex maximization problem. Through constructing an example network as an illustrative model of real

network conditions, we show, how complicated the underlying optimization problem multipath TCP may face, and through measurements, demonstrate how the various congestion mechanisms deal with finding a solution. The results appeared as a demo at ACM SIGCOMM 2019:
Zongor, L., Heszberger, Z., Pašić, A., & Tapolcai, J.: The Performance of Multi-Path TCP with Overlapping Paths, In Proceedings of the ACM SIGCOMM 2019 Conference Posters and Demos (pp. 116-118). ACM., 2019

In terms of the reliability of data transmission, we presented the deployment of a monitoring-flow based network verification and failure localization approach for SDN networks. It not only minimizes the number of static forwarding rules but also significantly reduces the control plane load, i.e., reduces the total number of messages needed for network verification and failure localization. Our flexible hybrid implementation consisted of MikroTik RB2011iLSIN and Open vSwitch (Mininet) switches, enabling the user to test network verification and failure localization in a more complex manner even if the number of physical devices is limited. Our results have been published at the 2019 INFOCOM conference:
Ladóczki Bence, Tapolcai János, Pašić, Alija: Monitoring-Flow Based Network Verification and Failure Localization in Software Defined Networks, In 38th IEEE International Conference on Computer Communications (INFOCOM), 2019

For supporting network reliability, we have presented a proof-of-concept design and implementation of an extensible automated failure localization framework. Designing, implementing, and maintaining network policies that protect from internal and external threats is a highly non-trivial task. Often, troubleshooting networks consisting of diverse entities realizing complex policies are even harder. Software-defined networking (SDN) enables networks to adapt to changing scenarios, which significantly lessens the human effort required for constant manual modifications of device configurations. Troubleshooting benefits SDN's method of accessing forwarding devices as well, since monitoring is made much easier via unified control channels. However, by making policy changes easier, the job of troubleshooting operators is made harder too: For humans, finding, analyzing, and fixing network issues becomes almost intractable. We presented a failure localization framework and its proof-of-concept prototype that helps in automating the investigation of network issues. Like a controller for troubleshooting tools, our framework integrates the formal specification (expected behavior) and network monitoring (actual behavior) and automatically gives hints about the location and type of network issues by comparing the two types of information. By using NetKAT (Kleene algebra with tests) for formal specification and Felix and SDN traceroute for network monitoring, we showed that the integration of these tools in a single framework can significantly ease the network troubleshooting process. The corresponding results have been published in a journal article in Future Internet:
István Pelle, András Gulyás: An Extensible Automated Failure Localization Framework Using NetKAT, Felix, and SDN Traceroute, FUTURE INTERNET 11 : 5 p. 107, 2019

We have also studied how to increase the availability of a backbone network with minimal cost. In particular, we presented a new resiliency framework focusing on resilience against natural disasters. It targets three different directions, namely: network planning, failure modeling, and survivable routing. The steady-state network planning is tackled by upgrading a sub-network (a set of links termed the spine) to achieve the targeted availability threshold. A new two-stage approach is proposed: a heuristic algorithm combined with a mixed-integer linear problem to optimize the availability upgrade cost. To tackle the disaster-resilient network planning problem, a new integer linear program is presented for the optimal link intensity tolerance upgrades together with an efficient heuristic scheme to reduce the running time. Failure modeling is improved by considering more realistic disasters. In particular, we focused on earthquakes using the historical data of the epicenters and the moment magnitudes. The joint failure probabilities of the multi-link failures were estimated, and the set of shared risk link groups is defined. The survivable routing aims to improve the network's connectivity during these shared risk link group failures. Here, a generalized dedicated protection algorithm was used to protect against all the listed failures. Finally, the experimental results demonstrated the benefits of the refined eFRADIR framework in the event of disasters by guaranteeing low disconnection probabilities even during large-scale natural disasters. Corresponding results have been published at the RNDM conference, IEEE Access journal, and INFOCOM 2021:

A. Pašić, R. Girao-Silva, B. Vass, T. Gomes, F. Mogyorósi, P. Babarczi, and J. Tapolcai: FRADIR-II: An Improved Framework for Disaster Resilience, Int. Workshop on Resilient Networks Design and Modeling (RNDM), Nicosia, Cyprus, 2019., 2019,

A. Pašić, R. Girão-Silva, F. Mogyorósi, B. Vass, T. Gomes, P. Babarczi, P. Revisnyei, J. Tapolcai, J. Rak: eFRADIR: An Enhanced FRAmework for DIsaster Resilience, IEEE Access (Volume: 9), 2021,

János Tapolcai; Zsombor L. Hajdú; Alija Pašić; Pin-Han Ho; Lajos Rónyai: On Network Topology Augmentation for Global Connectivity under Regional Failures IEEE INFOCOM 2021 - IEEE Conference on Computer Communications, 2021

A theoretical model for routing scalability

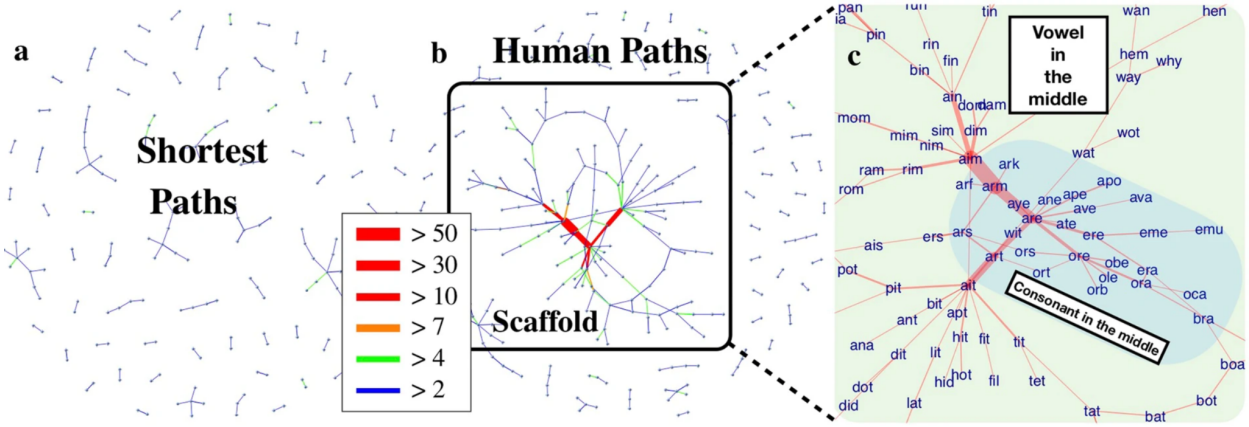
Routing in large-scale computer networks today is built on hop-by-hop routing: packet headers specify the destination address and routers use internal forwarding tables to map addresses to next-hop ports. In this project, we took a new look at the scalability of this paradigm. We defined a new model that reduces forwarding tables to sequential strings, which then lend themselves readily to an information-theoretical analysis. Contrary to previous work, our analysis is not of worst-case nature but gives verifiable and realizable memory requirement characterizations even when subjected to concrete topologies and routing policies. We formulated the optimal address space design problem as the task to set node addresses in order to minimize certain network-wide entropy-related measures. We derived tight space bounds for many well-known graph families and we propose a simple heuristic to find optimal address spaces for general graphs. Our evaluations suggest that in structured graphs, including most practically important

network topologies, significant memory savings can be attained by forwarding table compression over our optimized address spaces. According to our knowledge, our work is the first to bridge the gap between computer network scalability and information-theory:
A. Kőrösi, A. Gulyás, Z. Heszberger, J. Bíró and G. Rétvári: On the Memory Requirement of Hop-by-Hop Routing: Tight Bounds and Optimal Address Spaces, IEEE/ACM Transactions on Networking, vol. 28, no. 3, pp. 1353-1363, 2020

Besides information-theoretical analysis, we have also published improvements regarding the scalability of packet processing mechanisms in contemporary routers. Packet processing programs may have multiple semantically equivalent representations in terms of the match-action abstraction exposed by the underlying data plane. Some representations may encode the entire packet processing program into one large table allowing packets to be matched in a single lookup, while others may encode the same functionality decomposed into a pipeline of smaller match-action tables, maximizing modularity at the cost of increased lookup latency. We provided the first systematic study of match-action program representations in order to assist network programmers in navigating this vast design space. Borrowing from the relational database and formal language theory, we defined a framework for the equivalent transformation of match-action programs to obtain certain irredundant representations that we call “normal forms”. We found that normalization generally improves the capacity of the control plane to program the data-plane and to observe its state, at the same time having negligible, or positive, performance impact. Corresponding publications:
Felicián Németh, Marco Chiesa, and Gábor Rétvári: Normal forms for match-action programs, ACM CoNEXT, 2019
Balázs Vass, Erika Bérczi-Kovács, Costin Raiciu, Gábor Rétvári: Compiling Packet Programs to Reconfigurable Switches: Theory and Algorithms, EuroP4'20: Proceedings of the 3rd P4 Workshop in Europe December 2020 Pages 28–35, 2020

Study and modeling of human navigation schemes

Our main objective here was to identify the signal propagation and delivery paths, including the option for multipath, that arises in the course of human navigation over unknown network topologies and unassisted human learning, using a mobile application we have developed for an earlier study (fit-fat-cat). It is interesting, that despite their importance for public transportation, communication within organizations, or the general understanding of organized knowledge, our understanding of how human individuals navigate in complex networked systems is still limited owing to the lack of datasets recording a sufficient amount of navigation paths of individual humans. We analyzed 10587 paths recorded from 259 human subjects when navigating between nodes of a complex word-morph network. We identified the clear presence of systematic detours organized around individual hierarchical scaffolds guiding navigation. Our dataset was the first to enable the visualization and analysis of scaffold hierarchies (see figure below) whose presence and role in supporting human navigation is assumed in existing navigational models.



Showing the presence of scaffolds in the human navigation process on the three-letter word network

Detailed statistical evaluation was presented, showing that the presence of detours cannot be explained with randomized null models. In addition, we show that taking short detours following the hierarchical scaffolds is a clear sign of human subjects simplifying the interpretation of the complex networked system. To show this, we adapted our information-theoretic model, originally developed for computer networks (see above), to word networks and computed the entropy of shortest path and scaffold-aided path selection. We also discussed the role of these scaffolds in the phases of learning to navigate a network from scratch. These results were published at:

Gulyás, A., Bíró, J., Rétvári, G. et al.: The role of detours in individual human navigation patterns of complex networks, Nature Scientific Reports 10, 1098, 2020

András Gulyás: The role of detours in individual human navigation patterns of complex networks, Proc. of FENS Dynamics of the brain: temporal aspects of computation, Copenhagen, 2019

András Gulyás
Principal Investigator