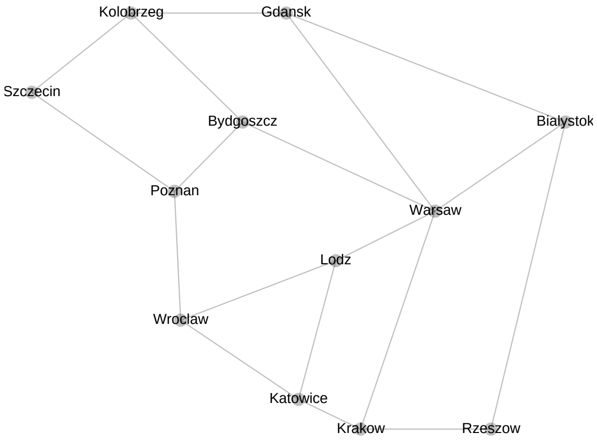
Since the inception of the Internet, the continuous growth of network traffic is being observed. The pace of this growth does not decrease. On the contrary, the shift of many activities to online form, imposed by the breakout of the COVID-19 pandemic, increased it further. To face up a growing demand, network operators have to expand their networks, which often means expensive infrastructural investments (e.g. laying a new cable between cities). Another solution allowing to manage rapidly increasing demand is the reduction of offered service quality. This approach was applied during the first lockdown when the European Commission imposed a reduction of default-offered video resolution on big VoD providers (YouTube, Netflix, etc.).

However, there are alternative solutions, which do not involve quality reductions or expensive investments. They come down to more intelligent traffic management in the existing infrastructure. The current state-of-the-art in networks is the usage of many available paths to send traffic between two particular nodes in a network, but only when their costs (lengths) are equal (so-called *ECMP*). This means that, for example, to send traffic between Krakow and Lodz two paths will be used: Krakow-Warsaw-Lodz and Krakow-Katowice-Lodz. However, to send traffic between Krakow and Warsaw, only one direct path will be used (Krakow-Warsaw), because there are no alternative paths of the same length. When this path becomes overloaded, alternative longer paths (e.g. Krakow-Rzeszow-Bialystok-Warsaw) will not be used, despite they may contain significant bandwidth reserves.



The solution to this problem is the usage of flow-based adaptive and multipath routing. This means a dynamic selection of paths for subsequent flows, taking into account the current or predicted load in the network. As the majority of traffic in networks is caused by a small number of large flows (so-called *elephant flows*), significant improvements can be achieved by implementing such individual per-flow routing only for this small group of flows.

For a few years this has been the most promising research direction in traffic engineering (TE). There are several dozen research articles published yearly, proposing complete flow-based TE systems, which authors claim to solve the aforementioned problems in the best possible way. However, the key variable affecting the results presented in these papers is the used traffic model, in particular the distribution of flow length and size. Relevant and credible models are not available in the literature. As a result, every several dozen TE systems proposed yearly is evaluated using different traffic models. This makes their comparison impossible and effectively blocks scientific and technological progress. Moreover, the used models are often oversimplified or deliberately cherry-picked to show the superiority of a particular system.

Having in mind the above, the goal of this project will be the creation of precise and reproducible statistical models of flows on the Internet. Created models will be able to be used both as input to analytical calculations and to generate realistic traffic in network simulators. The source of data will be CAIDA packet traces, MAWI packet traces, and own traces collected on the Internet-facing interface of a big campus network. Models will be created for general Internet traffic, as well as subsets of traffic (TCP-only, UDP-only, excluding DNS, etc.). Various flow definitions will be considered. They will include both network (host-to-host) and transport (5-tuple) layer flows and several flow timeout values, starting from milliseconds (so-called flowlets), up to 60 seconds. A reusable library of created models will be published online as open source. A user, after selecting the trace source and flow definition will be able to download the corresponding model. The final task will be an attempt to derive a single parameterized model which will be able to cover flows with various timeouts.