Large Scale Network Analytics with SNAP

Tutorial at the World Wide Web 2015 Conference

Rok Sosič
Stanford University
rok@cs.stanford.edu

Jure Leskovec
Stanford University
jure@cs.stanford.edu

ABSTRACT

Many techniques for the modeling, analysis and optimization of Web related datasets are based on studies of large scale networks, where a network can contain hundreds of millions of nodes and billions of edges. Network analysis tools must provide not only extensive functionality, but also high performance in processing these large networks.

The tutorial will present Stanford Network Analysis Platform (SNAP), a general purpose, high performance system for analysis and manipulation of large networks. SNAP is being used widely in studies of the Web datasets. SNAP consists of open source software, which provides a rich set of functions for performing network analytics, and a popular repository of publicly available real world network datasets. SNAP software APIs are available in Python and C++.

The tutorial will cover all aspects of SNAP, including APIs and datasets. The tutorial will include a hands-on component, where the participants will have the opportunity to use SNAP on their computers.

Categories and Subject Descriptors

H.4 [**Information System Applications**]: Data Mining.

Keywords

Graph analytics; graph processing; networks.

1. BACKGROUND

The emergence of the Web represents a fundamental shift as it has added important new dimensions to the production and dissemination of information. The volume of information available on the Web far exceeds what was available before the Web. The Web also replaces the traditional one-way mass-media to consumer communication channel with an interactive dialogue, which allows for the creation and exchange of user-generated content and provides a connection between our social networks, personal information channels and the mass media.

Web content in the form of organizational or user generated content on Websites, blogs, microblogs like Twitter, discussion forums, product review and multimedia sharing Websites presents many new opportunities and challenges. Companies and researchers analyze large scale networks to perform analytics, sentiment analysis or find influencers. Although there is a vast quantity of data available, the consequent challenge is to be able to analyze the large volumes of available content on the Web and often implicit links between the content, in order to gain meaningful insights.

These insights are gained using techniques for data modeling, analytics and optimization, based on studies of large scale networks. Network analysis tools must provide extensive functionality, while at the same offering high performance in processing these large networks with hundreds of millions of nodes and billions of edges.

2. DESCRIPTION

The tutorial will give an overview of basic principles of network analytics and how to use SNAP for network analytics in real world scenarios.

Stanford Network Analysis Platform (SNAP) is a general purpose, high performance system for analysis and manipulation of large networks. SNAP is being used widely in studies of Web-based and other large scale networks. SNAP consists of software, which provides a rich set of functions for performing network analytics and is available for Python and C++, and a popular repository of real world network datasets [1, 2]. All software is freely available under a liberal open source license. All datasets discussed are publicly available for download from the Web.

The tutorial is designed to proceed from entry level to more advanced topics. At the end of the tutorial, the participants will understand basics of network analytics, the resources provided by SNAP and how to apply those resources to network analytic tasks on Web-based datasets. They will have SNAP installed on their computers and will gain hands-on experience with SNAP. The tutorial is structured in 5 parts: Python API, C++ API, analytic functionality, datasets, hands-on exercises.

Complete tutorial materials are available at: `http://snap.stanford.edu/proj/snap-www`.

An earlier version of the tutorial was presented at ICWSM-14, Ann Arbor, MI in June 2014.

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2744708.

3. DETAILED OUTLINE

A more detailed outline of the tutorial is as follows:

- SNAP Python API
  - introduction, documentation, installation; basic types; vectors, hash tables, pairs; graph and network types; graph creation; graph traversal; graph saving and loading; graph manipulation.
- SNAP C++ API
  - introduction, documentation, installation; graph and network types; graph creation; graph traversal; graph saving and loading; graph manipulation; create your own project.
- SNAP analytic functionality
  - connected components; node degrees; breadth first search and depth first search based properties; node centrality; K-core decomposition; decompositions of graph adjacency matrix; counting triads; community detection; subgraphs; graph generators; plotting; graph visualization; advanced functionality.
- SNAP datasets
  - social networks; networks with ground-truth communities; communication networks; citation networks; collaboration networks; Web graphs; Amazon networks; Internet networks; road networks; autonomous systems; signed networks; location-based online social networks; wikipedia networks and metadata; Twitter and Memetracker; online communities; online reviews.
- hands-on session
  - SNAP installation on participants computers
  - hands-on exercises

4. TARGET AUDIENCE

The intended audience for the tutorial are participants that want to learn principles of network analytics and how to apply them in real world scenarios. The tutorial is targeted toward entry level audience with some programming background, thus the Python API will be presented in more detail than the C++ API. Some advanced topic materials will be covered as well.

The goal of the tutorial is for the participants to learn how to apply network analytic methods to practical problems. The participants will install SNAP on their computers and gain hands-on experience with SNAP through a set of exercises.

Participants are expected to have entry level programming experience, preferably with Python or C++ for maximum benefit. Background in basic graph and network concepts and algorithms is helpful although it is not required. Tutorial materials will be provided to all attendees and will be freely accessible via the Web after the tutorial.

To gain hands-on experience, the participants working with Python must have a computer with one of the following configurations: 64-bit MS Windows with Visual Studio (a free version is available from Microsoft) and Python 2.x, 64-bit Mac OS X with Python 2.x, 64-bit Linux with Python 2.x. For participants that want to use C++, a working C++ compiler and associated development tools are required.

5. PRESENTERS

Rok Sosič. Rok (rok@cs.stanford.edu) is a senior researcher in Prof. Leskovec's group at Stanford University, working on tools for large scale network analytics. He published over 40 papers, including the best paper at Supercomputing'95 and a top 10 paper in the field of high-performance distributed computing (www.hpdc.org/best.php). He lead one of the first grid computing deployments on Wall Street and later headed engineering at Turbolinux, a world's top 3 Linux distribution at that time with 10's of millions of users. Most recently, he lead engineering efforts at SkyGrid with the App Store Best of 2011 News App. Rok received his PhD in Computer Science from University of Utah. He joined Stanford University in 2012.

Jure Leskovec. Jure (jure@cs.stanford.edu) is assistant professor of Computer Science at Stanford University. His research focuses on mining large social and information networks. Problems he investigates are motivated by large scale data, the Web and on-line media. This research has won several awards including a Microsoft Research Faculty Fellowship, the Alfred P. Sloan Fellowship and numerous best paper awards. Leskovec received his bachelor's degree in computer science from University of Ljubljana, Slovenia, and his PhD in in machine learning from the Carnegie Mellon University and postdoctoral training at Cornell University.

6. REFERENCES

[1] Jure Leskovec and Andrej Krevl. SNAP Datasets: Stanford large network dataset collection. `http://snap.stanford.edu/data`, June 2014.

[2] Jure Leskovec and Rok Sosič. SNAP: A general purpose network analysis and graph mining library in C++. `http://snap.stanford.edu/snap`, June 2014.