CPTS 415
Resources and Datasets

Instructor: Yinghui Wu
EME B45, School of EECS, WSU
yinghui@eecs.wsu.edu

Datasets: You are encouraged to go over and understand the public datasets below; think about interesting problems and come up with your own project, leveraging the techniques, design principles and tools we introduced in the Big Data course.

- Stanford Large Network Dataset Collection
  60 large social and information network datasets

- Coauthorship and Citation Networks
  DBLP: Collaboration network of computer scientists
  KDD Cup Dataset

- Internet Topology
  AS Graphs: AS-level connectivity inferred from Oregon route-views, Looking glass data and Routing registry data
  CAIDA data: AS-level network traffic data

- Stack Overflow Stack Overflow Data

- Yelp Data Yelp Review Data: reviews of the 250 closest businesses for 30 universities for students and academics to explore and research

- Prosper peer to peer money lending dataset
  Money Lending Data: Lenders ask for loans and people bid (price, interest rate) on loans to fund.

- Youtube dataset
  Youtube data: YouTube videos as nodes. Edge a->b means video b is in the related video list (first 20 only) of a video a.

- Amazon product copurchasing networks and metadata
  Amazon Data: The data was collected by crawling Amazon website and contains product metadata and review information about 548,552 different products (Books, music CDs, DVDs and VHS video tapes).

- Wikipedia
  Wikipedia page to page link data: A list of all page-to-page links in Wikipedia

- DBPedia: The DBpedia data set uses a large multi-domain ontology which has been derived from Wikipedia.

- Edits and talks: Complete edit history (all revisions, all pages) of Wikipedia since its inception till January 2008.

- Movie Ratings
  IMDB database: Movie ratings from IMDB
  User rating data: Movie ratings from MovieLens

- Who trusts whom data at Trustlet
  Trust network datasets: Includes trust/distrust edges and Epinions product reviews/review ratings

- Mark Newman's pointers
  Network data: More than 20 network datasets

- Munmun De Choudhury's pointers
  Network data: Flickr Image Dataset, YouTube Dataset, Digg Dataset (Social Media), Engadget Dataset (online communities), Del.icio.us Dataset (Social bookmarking)

- Reality Commons data
  Mobile data: Several mobile data sets that contain the dynamics of several communities of about 100 people each.

- Global Terrorist Data A table of global terrorist events, location, organization, etc.

E-Books: Feel free to explore the online e-books below as extended reading related with the course. These books cover from basic Big Data concepts and tools to the state-of-the-art Big Data industry applications, management and cultural demand of being data driven.

- Big data for dummies
  Big Data For Dummies cuts through the confusion and helps you take charge of big data solutions for your organization.
- Big Data Glossary
  To help you navigate the large number of new data tools available, this guide describes 60 of the most recent innovations, from NoSQL databases and MapReduce approaches to machine learning and visualization tools. Descriptions are based on first-hand experience with these tools in a production environment.
- Big Data Now
  The topics in this 2014 edition of Big Data Now represent the major forces currently shaping the data world.
- Hadoop: The definitive Guide *Popular Hadoop tutorial*
- Building Data Science Teams
  Data science teams need people with the skills and curiosity to ask the big questions.
- Planning for Big Data
  As this emerging field transitions from the bleeding edge to enterprise infrastructure, it's

vital to understand not only the technologies involved, but the organizational and cultural demands of being data-driven.

- Statistics for Machine Learning *Statistics 101 for Machine learning*
- What is Data Science
  This report examines the many sides of data science -- the technologies, the companies and the unique skill sets.
- Mining of Massive Dataset *The book is based on Stanford Computer Science course CS246: Mining Massive Datasets by Jure Leskovec et.al*
- Data Jujitsu *The art of turning data into product*
- Ethics of Big Data *How to balance risk and innovation?*
- The intelligent Web good entry point for search and smart algorithms over big data, from the perspective of Web search engines.

Tools: Get familiar with the following open source tools and platforms, and select one or two that you will use for the course project.

MapReduce platforms/noSQL databases

- Hadoop: https://hadoop.apache.org/releases.html

- MongoDB: leading noSQL database: https://www.mongodb.com/leading-nosql-database
- CouchDB: a database for Web http://couchdb.apache.org/
- Apache Spark: fast and general large-scale data engine http://spark.apache.org/

Graph databases and engines

- GraphLab: popular distributed/parallel graph analytics https://dato.com/ http://select.cs.cmu.edu/code/graphlab/download.html
- Neo4j: popular, light weighted graph database http://neo4j.com/download/

Open source graph visualization tools

- Gephi: popular graph visualization tool http://gephi.github.io/
- D3: data driven document visualization http://d3js.org/
- Cytoscape: complex network integration and visualization http://www.cytoscape.org/

Graph library/Algorithm set

- JGraphT: an open Java graph library
  http://jgrapht.org/
- JUNG: Java universal network/graph framework
  http://jung.sourceforge.net/

Related reading:

http://www.itbusinessedge.com/slideshows/top-five-nosql-databases-and-when-to-use-them.html

System Installation

Below we briefly introduce the installation of several platforms on your machine (thanks for Qi's work!). You should figure out how to execute a sequential algorithm in a "pseudo-distributed" environment using e.g., Hadoop.

Single-node cluster for Hadoop
Single-node cluster for GraphLab
Single-node cluster for Spark