Project Topics

Below is a list of possible project topics. Some of these are open-ended, meaning that you are required to come up with a new algorithm or model, and formulate it yourselves. Such projects may require more effort, but they will be also graded based on the effort, as well as the final result. Others are more straight-forward, you would need to obtain a complex dataset and apply algorithms on this data. There are also more theoretical projects, and more practical ones, so you can pick depending on your preference.

You will also have to present in class one paper related with you project. The list below includes the paper for each project.

Projects should be done **in teams of at most two** students.

Deliverables and Timeline:

- A web page with all information related to your project (first version: week before Christmas, final version: end of January)
- A two- page project proposal outlining what you plan to do. This should include the topic of your presentation (first version: week before Christmas)
- A 20' presentation of 1-2 research papers related to your project (presentations are scheduled for 15/1 and 22/1)
- The source code of your project (end of January)
- A final report describing your project following a specific format (end of January)

Topic 1

GraphChi is a disk-based system that supports efficient processing of very large graphs.
http://graphlab.org/graphchi/

Project:
In this project, you need first to install GraphChi and then, implement a number of centrality measures (e.g., degree, distance, Pagerank). The implementation of most of these measures may be already available in GraphChi.
Then, you will use them to design, implement and evaluate appropriate heuristics for the following problem: Given two graph snapshots that correspond to a social graph at two different time instants, identify the *k* pair of nodes whose shortest path distance has changed the most. Intuitively, those are the pairs of users who came closer to each other. For example, a possible heuristic would be to consider the pairs (u, v) where u or v is among the m < k nodes whose degree, PageRank etc has increased the most.

Paper:
A. Kyrola, G. Blelloch and C. Guestrin, GraphChi: large-scale graph computation on just a PC, OSI 2012

Topic 2

Pregel is a distributed framework developed at Google for processing large graphs. It follows a bulj synchronous parallel (BSP) model where vertices send messages to other vertices in supersteps.

Giraph is an open source implementation of Pregel (runs on standard Hadoop infrstastruture).
http://graphlab.org/graphchi/

GPS is another implementation of Pregel available as open source under the BSD license.
http://infolab.stanford.edu/gps/

Project:
In this project, you need first to install one of the two implementations of Pregel.
Then, implement at least two different algorithms for computing betweness and evaluate their performance. For the evaluation, you should use at least three different graphs (real or synthetic) and different system configurations.

Paper:
Grzegorz Malewicz, Matthew H. Austern, Aart J. C. Bik, James C. Dehnert, Ilan Horn, Naty Leiser, Grzegorz Czajkowski: Pregel: a system for large-scale graph processing. SIGMOD Conference 2010: 135-146

Topic 3

Foursquare is an online location-based social network where users can check-in in venues and leave tips.

Project:
First, use the FourSquare API
https://developer.foursquare.com/
to collect data.
Then, perform analysis on the collected datasets along the lines of the WSDM12 paper below to confirm, disprove or extend its findings.

Paper:
Marisa A. Vasconcelos, Saulo M. R. Ricci, Jussara M. Almeida, Fabrício Benevenuto, Virgílio A. F. Almeida: Tips, dones and todos: uncovering user profiles in foursquare. WSDM 2012: 653-662

Topic 4

Project:
Download the Facebook data related to a small number (2-5) users.

Then, use lucene http://lucene.apache.org/, a free, open source information retrieval software library to index and search the collected data. The project can be extended to add PageRank in the network of friends to improve ranking.

Paper:
Chia-Jung Lee, W. Bruce Croft, Jinyoung Kim: Evaluating search in personal social media collections. WSDM 2012: 683-692

Topic 5

Project:
Use the Twitter API
https://dev.twitter.com/
to collect datasets from Twitter to extend the analysis in
Konstantinos Semertzidis, Evaggelia Pitoura, Panayiotis Tsaparas: How people describe themselves on Twitter. DBSocial 2013: 25-30
In particular, you should collect information about the user profile (bio), analyze it and then test whether it can be used for link prediction.

Paper:
John E. Hopcroft, Tiancheng Lou, Jie Tang: Who will follow you back?: reciprocal relationship prediction. CIKM 2011: 1137-1146

Topic 6

Project:
Use the Twitter API
https://dev.twitter.com/
in particular, the search API to collect tweets related to at least 5 well-known events (e.g., Philippines typhoon) and then analyze how the related hashtags evolve over time.

Paper:
Farshad Kooti, Haeryun Yang, Meeyoung Cha, P. Krishna Gummadi, Winter A. Mason: The Emergence of Conventions in Online Social Networks. ICWSM 2012

Topic 7

Use the Flickr API
http://www.flickr.com/services/api/
to collect datasets from Flickr.
Then perform some analysis on the collected datasets.
Potential topics: identifications of POIs (points of interest), etc

Paper:
Tye Rattenbury, Nathaniel Good, Mor Naaman: Towards automatic extraction of event and place semantics from flickr tags. SIGIR 2007: 103-110

Topic 8

Project:

Use the GitHub API
http://developer.github.com/v3/
to collect datasets from GitHub.
Then perform some analysis on the collected datasets.
Potential topics: perform PageRank, team formation, expert location etc
The following paper
Laura A. Dabbish, H. Colleen Stuart, Jason Tsay, James D. Herbsleb: Social coding in GitHub: transparency and collaboration in an open software repository. CSCW 2012: 1277-1286
contains some information about Github.

Paper:

Alan Mislove, Massimiliano Marcon, P. Krishna Gummadi, Peter Druschel, Bobby Bhattacharjee: Measurement and analysis of online social networks. Internet Measurement Comference 2007: 29-42

Topic 9

Graph Similarity

Project:

Implement the various similarity measures proposed in the paper below. Propose an extension that takes into account edge and/or node labels. Report evaluation results of applying the similarity measures on various graph datasets.

Paper:

Christos Faloutsos, Danai Koutra, Joshua T. Vogelstein: DELTACON: A Principled Massive-Graph Similarity Function. SDM 2013: 162-170

Topic 10

Sampling of Graphs

Project:

Propose a method for sampling a graph such that we can measure different properties? For example betweeness?

Example paper: Jure Leskovec, Christos Faloutsos *Sampling from Large Graphs* (poster) KDD 2006, Philadelphia, PA.

Topic 11

Finding important edges in time-evolving graphs

Project:

Given two graph snapshots that correspond to a social graph at two different time instants, which are the k edges whose addition affected the largest number of shortest paths?
The goal is to develop and test a number of efficient heuristics for identify such edges.

Paper:

Hanghang Tong, Spiros Papadimitriou, Philip S. Yu, Christos Faloutsos: Fast Monitoring Proximity and Centrality on Time-evolving Bipartite Graphs. Statistical Analysis and Data Mining 1(3): 142-156 (2008)

Topic 12

Team formation

Project:

Extend the model and algorithms for team formation described in the paper below for the case of negative edges. Implement and test them.

Paper:

Theodoros Lappas, Kun Liu, Evimaria Terzi: Finding a team of experts in social networks. KDD 2009: 467-476

Topic 13

Another option is to suggest a project of your own, based on what you have seen in the class so far, questions you may have thought of, and things that are related to your research area. In this case you should create a project proposal (initially just a paragraph or an idea) and contact us to discuss it.