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.

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.

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.

Timeline:

- Week before Christmas: Submit a ~2-page project proposal outlining what you plan to do. This should include the topic of your presentation
- First two weeks after Christmas: Presentations.
    - Present one or more papers or background material related to your project
- End of January: Submit full project.

Topic 1

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

Giraph is an open source implementation of Pregel (runs on standard Hadoop infrastructure) http://incubator.apache.org/giraph/

Project:

Install Pregel. Implement PageRank (and/or Betweeness) on a synthetic (power law) and a real graph. Results of a performance evaluation for different configurations and graph parameters should also be reported.

Paper:

G. Malewicz, M. H. Austern, A. J. C. Bik, J.C. Dehnert, I. Horn, N. Leiser, G. Czajkowski: *Pregel: a system for large-scale graph processing*. SIGMOD Conference 2010: 135-146

Team: Χ. Θεοδωράκης Α. Παπαδογιαννάκης

Topic 2

MapReduce is a distributed framework developed at Google for processing big data.

Hadoop is an open source implementation of MapReduce
http://hadoop.apache.org/

Project:

Install MapReduce. Implement PageRank (Betweness) on a synthetic (power law) and a real graph. Results of a performance evaluation for different configurations and graph parameters should also be reported.

Paper:

Graph Twiddling in MapReduce as a starting point.

Useful paper

J. Lin and M. Schatz *Design Patterns for Efficient Graph Algorithms in MapReduce*, http://www.umiacs.umd.edu/~jimmylin/publications/Lin_Schatz_MLG2010.pdf

Team: Θ. Βαρτζιώτης Χ. Γιαννάκος

Topic 3

Project:

Use the FourSquare API
https://developer.foursquare.com/
to collect data.
Then, perform analysis on the collected datasets. Potential topics: compare the social graph among different cities/countries; study homophily, create personal trajectories, etc

Paper:

J. Lindqvist, J. Cranshaw, J. Wiese, J. I. Hong, J. Zimmerman: *I'm the mayor of my house: examining why people use foursquare - a social-driven location sharing application*. CHI 2011: 2409-2418
A. Noulas, S. Scellato, C.Mascolo, M. Pontil: *An Empirical Study of Geographic User Activity Patterns in Foursquare*. ICWSM 2011

Team: Μ. Κονταξή Ι. Κοτρώτσιος

Topic 4

Project:

Use the Facebook API
http://developers.facebook.com/docs/reference/api/
to collect datasets from Facebook.

Then, perform analysis on the collected datasets. Potential topics: information cascading, identify spammers, homophily between friends, patterns in friends that comment on posts, determining relationship strength.

Paper:
B. Viswanath, A. Mislove, M. Cha, P. Krishna Gummadi: *On the evolution of user interaction in Facebook.* WOSN 2009: 37-42

Team: Ε. Γατσιώρη Χ. Κλιτσινάρη

Topic 5

Project:
Use the Twitter API
https://dev.twitter.com/
to collect datasets from Twitter.
Then, perform some analysis on the collected datasets.
Potential topics: expert identification, spammers, similarity between friends, hashtags used between friends.

Paper:
H. Kwak, C. Lee, H. Park, S. B. Moon: *What is Twitter, a social network or a news media?* WWW 2010: 591-600

Team: Κ. Καλτσάς Γ. Παπαγιάννης

Topic 6

Project:
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:
M. Cha, A.Mislove, P. K. Gummadi: *A measurement-driven analysis of information propagation in the flickr social network*. WWW 2009: 721-730

Topic 7

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, etc

Paper:

A. Mislove, M. Marcon, P. K. Gummadi, P.Druschel, B. Bhattacharjee: *Measurement and analysis of online social networks.* Internet Measurement Conference 2007: 29-42

Team: Α. Βόγλης

Topic 8

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:

M. Berlingerio, D. Koutra, T. Eliassi-Rad, C. Faloutsos*: NetSimile: A Scalable Approach to Size-Independent Network Similarity*. CoRR abs/1209.2684 (2012)

Topic 9

Project:

Implement some of the algorithms for team formation described in the paper below. Apply them on some real datasets and compare their performance. Alternatively, consider how they can be extended in the case of negative edges

Paper:

T. Lappas , K. Liu, E. Terzi. A Survey of Algorithms and Systems for Expert Location in Social Networks

http://link.springer.com/chapter/10.1007%2F978-1-4419-8462-3_8?LI=true

Team: Μ. Αργύρη Π. Ζαγορίσιος

Topic 10

Network Models:

Experiment with a new model for network generation

Possible ideas:

- Geometric models with copying of locations
- Using SOC (Self-Organized Criticality) for network modeling
- Affiliation networks.

The papers upon which the model is built.

E.g., affiliation networks paper: Lattazani, Sivakumar, “Affiliation Networks”, STOC 2009 (and follow-up on WWW 2010)

Team: Β. Παπαδόπουλος Ι. Χιόνης

Topic 11

Sampling of Graphs

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

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

Topic 12

Prediction of Friend and Enemy relationships: Give an algorithm that predicts if a link will appear, but also the sign of the link.

Paper: Predicting Positive and Negative Links in Online Social Networks by J. Leskovec, D. Huttenlocher, J. Kleinberg. *ACM WWW International conference on World Wide Web (WWW)*, 2010.

Team: Α. Φραγκούλης, Σ. Πολενάκης

Topic 13

Consider a cyclical sequence of graphs $G_1, \dots, G_T$ over the same nodes capturing the relationships between nodes as they evolve periodically over the course of a week. Find an algorithm for the following problem(s):

1. Which nodes to infect at G_1 so as to maximize the spread in the network using the independent cascade model?
2. Which nodes to infect at which time so as to maximize the spread in the network assuming the independent cascade model.

Possible dataset: Reality Mining project, data from the paper below

Paper: Cho, Myers, Leskovec, Friendship and Mobility: User Movement In Location-Based Social Networks, KDD 2011

David Kempe, Jon Kleinberg, and Amit Kumar. Connectivity and inference problems for temporal networks. In *Proc. 32nd ACM Symposium on Theory of Computing*, pages 504–513, 2000

Team: Ν. Γκαιρώ Ε. Πασχάλη