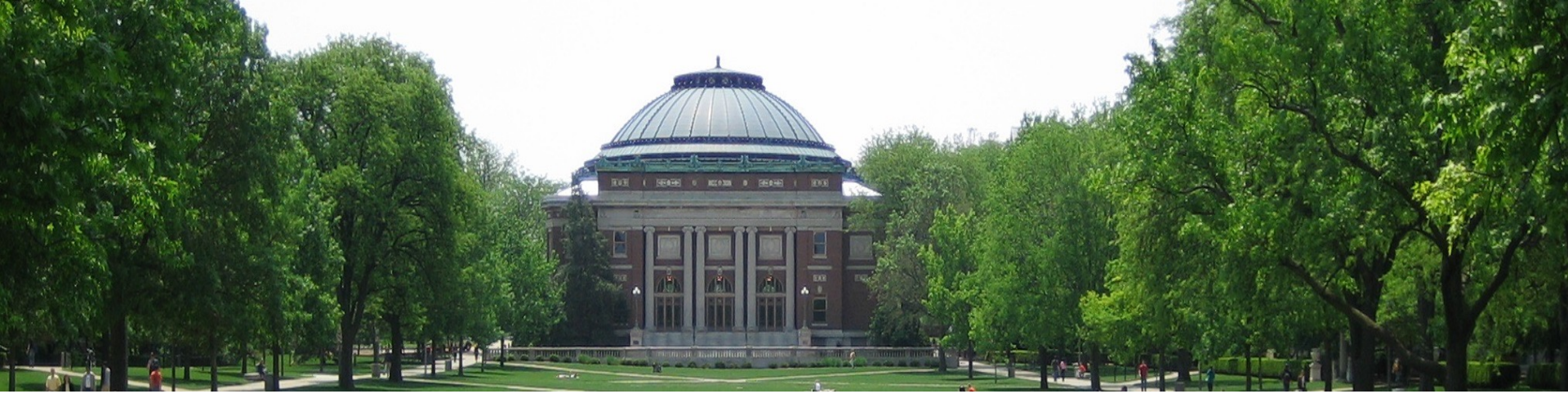


CS 512 Data Mining Principles

Class Project Guidance

Hanghang Tong, Computer Science, Univ. Illinois at Urbana-Champaign, 2022

Grading and Key Milestones

- **Class Project accounts for 30% of overall grade**
  - Project proposal 2% -- **Sep. 13th, 2022**
  - Mid-term report 8% -- **Oct. 18th, 2022**
  - Final report 20% -- **Dec. 12th, 2022**
  - All the time is in CT zone
- Individual project or group (up to 3 members) project
- For all the writings (proposal, mid-term report and final report), use ACM conference template (https://www.acm.org/publications/proceedings-template)
  - Violation to the above formatting might lead to a 20% penalty
  - All submissions will be done on Canvas.
  - No late submission or emails submissions will be accepted
  - No hand-written submissions will be accepted

Class Project

- **Project proposal (1-2 pages) due on Sep. 13th, 2022**
  - project title
  - team members: roles of each member
  - description of the problem you try to address
  - preliminary plan (milestones)
  - paper list
- **Mid-term report (3-4 pages) due on Oct. 18th, 2022**
  - Content from proposal
  - Current progress of the project
  - Plan of the rest milestones
- **Final project report (8-12 pages) due on Dec. 12th, 2022**
  - Think of your final report as a submission to a top-tier conference (e.g., The Web Conference), and we will grade your report by acting as a reviewer.
  - Six to ten pages for the body of the paper (including all figures, statistics, etc.),
  - plus up to one additional page with references that do not fit within the body pages,
  - plus one page to describe the contributions (roles + percentage) of each team members .

Programming Language

- Your call (c/c++, java, matlab, r, python….)
- Usage of 3rd-party codes
  - Totally fine to build certain parts of your system - but check the permission/license first.
  - A certain amount of coding from each team is expected.
  - Clearly state in your final report which parts are written by you and which are from the 3rd-party; and clearly state that you have the full permission of using such codes.
  - Do NOT submit the source code of the 3rd-party codes
  - You will not get credit for your grade for using 3rd-party codes alone. But if that helps your overall system, your overall system might get credit.
- For the part of codes that you want to claim credits for class projects, you must submit the source codes, together with your final report
- A significant amount of **your own code** is expected.

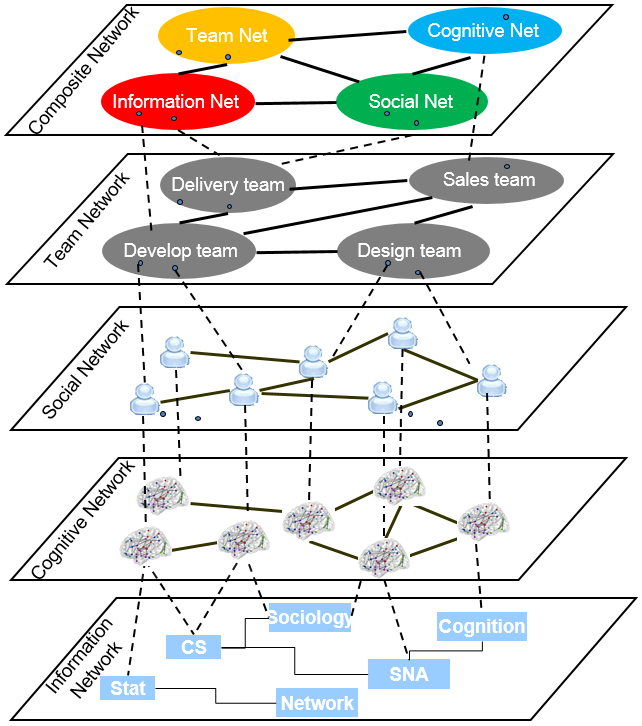
Candidate Projects

- Feel free to pick the projects outside the lists,
  - It must address certain data mining problems.
  - It must be tested on certain real data sets (i.e., purely testing it on synthetic data sets is not acceptable)
  - Whether or not choosing a project from the candidate list will **NOT** affect your grades (one way or the other)
- You ***CANNOT*** pick a class project from an existing data mining competition task (e.g., kaggle)

Project Ideas @ IDEA Lab
(2022)

Hanghang Tong
htong@illinois.edu

P1: Mining Network of Networks



- ***Problem***: Given a set of networks (e.g., social network from facebook, twitter social networks, etc), how to model them and find interesting patterns.
- ***Data***: Refer to introductory papers

- ***Introductory paper(s)***:
  - Jingchao Ni, Hanghang Tong, Wei Fan, Xiang Zhang: Inside the atoms: ranking on a network of networks. KDD 2014: 1356-1365
  - Chen Chen, Jingrui He, Nadya Bliss, Hanghang Tong: Towards Optimal Connectivity on Multi-Layered Networks. IEEE Trans. Knowl. Data Eng. 29(10): 2332-2346 (2017)
  - Multi-layered network embedding J Li, C Chen, H Tong, H Liu. SDM 2018
- ***Comments***: Very hot topics in web/network science in the recent years. Many possible extensions. Well likely to lead to publications.

P2: Spatial-Temporal Graph Mining

- ***Problem***: Spatial-temporal graphs appear in many applications, such as smart transportation, financial analysis, environmental monitoring and computer vision. Representative challenges include how to forecast the future values; impute the missing values; find anomaly patterns etc.
- ***Data****:* Many possible data sets, see the data sets in papers below.
- ***Introductory paper(s)*:**
  - Li Y, Yu R, Shahabi C, et al. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. ICLR, 2018.
  - Yan S, Xiong Y, Lin D. Spatial temporal graph convolutional networks for skeleton-based action recognition Thirty-second AAAI conference on artificial intelligence. 2018.
  - Atluri G, Karpatne A, Kumar V. Spatio-temporal data mining: A survey of problems and methods. ACM Computing Surveys (CSUR), 2018
- ***Comments***: open-ended, very broad applicability, might lead to publication.

P3: Understanding Network Mining: Attacking

- **Problem:**
  - In recent years, robustness on automated decision-making algorithms and models has drawn a lot of research attentions. Researchers try to find effective and efficient ways to attack those models in order to design better defensive strategies. Despite numerous research works on attacking image and text data, few attention has been paid to robustness on graph data. Choose a well-known graph mining task (e.g., ranking, clustering, classification), can you develop your method to effectively attack the corresponding model?
- **Datasets:**
  - SNAP (https://snap.stanford.edu/data/index.html)
- **Introductory materials:**
  - Hanjun Dai, Hui Li, Tian Tian, Xin Huang, Lin Wang, Jun Zhu, and Le Song. 2018. Adversarial attack on graph structured data. ICML 2018.
  - Daniel Zügner, Amir Akbarnejad, and Stephan Günnemann. 2018. Adversarial Attacks on Neural Networks for Graph Data. KDD 2018.
  - Qinghai Zhou, Liangyue Li, Nan Cao, Lei Ying, and Hanghang Tong: Admiring: Adversarial Multi-Network Mining. ICDM 2019
- **Comment:**
  - High impact research. Very hot topics in recent years. Well likely to lead to publication

P4: Embedding-based Network Alignment

- **Problem:**
  - In the era of big data, networks are often multi-sourced (i.e., variety of 5Vs). How to link networks from multiple sources has become a very important task for many applications, such as social analysis, recommendation, etc. Network alignment is to find node correspondence across networks and thus connect different networks together. Although many embedding based network alignment methods have been proposed, most of them bear the space disparity issue, which means that the embeddings of nodes in two networks usually fall into two different spaces. In that case, the node embeddings from two networks cannot be compared, which undermines the algorithms' performance. Can you propose some solution to solve this problem?
- **Datasets:**
  - SNAP network datasets (https://snap.stanford.edu/data/)
- **Introductory materials:**
  - Zhang, Si, and Hanghang Tong. "Final: Fast attributed network alignment." *KDD*. ACM, 2016.
  - Yuchen Yan, Si Zhang, Hanghang Tong. BRIGHT: A Bridging Algorithm for Network Alignment. ACM The Web Conference (WWW), 2021.
  - Si Zhang, Hanghang Tong, Long Jin, Yinglong Xia, Yunsong Guo. Balancing Consistency and Disparity in Network Alignment. ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2021.
  - Si Zhang, Hanghang Tong,, Yinglong Xia, Liang Xiong and Jiejun Xu. NetTrans: Neural Cross-Network Transformation. ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2020.
- **Comment:**
  - An important real-world problem, may lead to publications.

P5: Oversmoothing issue in GCN

- ***Problem****:* Oversmoothing is a serious issue for deep graph convolution network (GCN), can you analyze this issue and propose some solutions to solve it?
- ***Data****:* Many possible data sets, see the data sets in papers below.
- ***Introductory paper(s)***:
  - Chen Lei, Le Wu, Richang Hong, Kun Zhang, and Meng Wang. "Revisiting graph based collaborative filtering: A linear residual graph convolutional network approach." In *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 01, pp. 27-34. 2020.
  - Rong, Yu, Wenbing Huang, Tingyang Xu, and Junzhou Huang. "Dropedge: Towards deep graph convolutional networks on node classification." *arXiv preprint arXiv:1907.10903* (2019).
  - Li, Guohao, Matthias Muller, Ali Thabet, and Bernard Ghanem. "Deepgcns: Can gcns go as deep as cnns?." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9267-9276. 2019.
- ***Comments***: open-ended, very broad applicability, might lead to publication.

P6: Self-Supervised Learning

- ***Problem****:* Labels are the keys for training neural networks, which are usually expensive and time consuming to obtain in practice. Self-supervised learning aims to learn informative graph representations or graph encoders without labels.
- ***Data****:* Many possible data sets, see the data sets in papers below.
- ***Introductory paper(s)***:
  - Liu, Yixin, et al. "Graph self-supervised learning: A survey." arXiv preprint arXiv:2103.00111 (2021).
  - Liu, Xiao, et al. "Self-supervised learning: Generative or contrastive." IEEE Transactions on Knowledge and Data Engineering (2021).
  - Wu L, Lin H, Tan C, et al. Self-supervised Learning on Graphs: Contrastive, Generative, or Predictive[J]. IEEE Transactions on Knowledge and Data Engineering, 2021.
- ***Comments***: open-ended, very broad applicability, might lead to publication.

P7. Learning to Learn: Meta-Learning on Graphs

- **Problem:** effective supervised machine learning models normally requires a large amount of training data, however, humans can learn skills much faster and easily apply the learned knowledge to new environments. Meta-learning aims to learn efficiently with limited training data and to achieve fast generalization to new tasks, which is referred as "Learning to Learn". For example, on node classification task, we hope to achieve accurate classification with (1) very few labels from single source graph, or (2) labeled data from multiple graphs as the source.
- **Introductory papers:**
  - Liu, L., Zhou, T., Long, G., Jiang, J. and Zhang, C., 2019. Learning to propagate for graph meta-learning. arXiv preprint arXiv:1909.05024.
  - Huang, K. and Zitnik, M., 2020. Graph meta learning via local subgraphs. Advances in Neural Information Processing Systems, 33.
  - Finn, C., Abbeel, P. and Levine, S., 2017, July. Model-agnostic meta-learning for fast adaptation of deep networks. In International Conference on Machine Learning (pp. 1126-1135). PMLR.
- **Comments:** meta-learning is also related to few-shot learning and active learning, and is widely investigated in data mining research. It might lead to publication due to its broad applicability.

P8. Fairness in Graph Neural Networks

- **Problem:** Despite the recent success of Graph neural networks (GNNs) in many downstream tasks like node classification, link prediction and graph classification, evidence reveal that GNNs are often biased towards certain demographic groups defined by a sensitive attribute (e.g., gender, race). In this project, we would like to (1) ensure fairness in GNNs so that predictions made by GNNs are invariant to the sensitive attribute (2) without making much sacrifice on the overall performance of GNNs.
- **Introductory papers:**
  - Survey for GNNs: https://bit.ly/3ILyv84
  - Tutorial for fairness: https://bit.ly/3oMhNx9
  - Dai, E., & Wang, S.. Say no to the discrimination: Learning fair graph neural networks with limited sensitive attribute information. In WSDM 2021.
  - Bose, A., & Hamilton, W.. Compositional fairness constraints for graph embeddings. In ICML 2019.
- **Data:** See datasets in introductory papers
- **Comments:** Very hot research topic in recent years, high societal impact

P9. Active Learning on Knowledge Graphs

- **Problem:** The performance of knowledge graph models highly depends on the number of labeled samples. Active learning aims to identify samples that are most informative for training ML models. With effective active learning algorithms proposed for knowledge graphs, the annotation cost of knowledge graph tasks such as knowledge graph completion could be significantly reduced.
- ***Data****:* Many possible data sets, see the data sets in papers below.
- **Introductory papers:**
  - Ostapuk N, Yang J, Cudré-Mauroux P. Activelink: deep active learning for link prediction in knowledge graphs[C]//The World Wide Web Conference. 2019: 1398-1408.
  - Hu S, Xiong Z, Qu M, et al. Graph policy network for transferable active learning on graphs[J]. arXiv preprint arXiv:2006.13463, 2020.
  - Aggarwal, Charu C., et al. "Active learning: A survey." *Data Classification: Algorithms and Applications*. CRC Press, 2014. 571-605.
- **Comments:** Existing active learning works focus on homogeneous graphs, few of them study more complicated graph-structured data such as knowledge graphs. This project might lead to publication.

P10. Graph Mining beyond Homophily

- **Problem:** The homophily assumption is intuitive and widely-used in many graph mining tasks which assumes the connected nodes tend to share the same label/property. For example, node classification algorithms tend to smooth the representation of connected nodes, clustering methods tend to cluster highly-connected nodes into a group, and many more. Recent years, increasing attention is paid onto the graphs beyond homophily, on which connected nodes are less likely to have the same labels, such as the dating network. However, recent works are concentrated on the node classification tasks, especially methods equipped with graph neural nets. In fact, many existing graph mining algorithms may get into trouble on graphs beyond homophily and those problems are under-explored so far.
- ***Data****:* Refer to the reference papers.
- **Introductory papers:**
  - Zhu, Jiong, et al. "Beyond Homophily in Graph Neural Networks: Current Limitations and Effective Designs." Advances in Neural Information Processing Systems 33 (2020).
  - Chien, Eli, et al. "Adaptive universal generalized pagerank graph neural network." International Conference on Learning Representations. https://openreview. net/forum. 2021.
  - Bo, Deyu, et al. "Beyond Low-frequency Information in Graph Convolutional Networks." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 35. No. 5. 2021.
- **Comments:** Very competitive topic, but mostly on the node classification tasks. Can lead into interesting problem for many graph mining tasks. Remark: this problem is closely-related with over-smoothing problem in a high-level sense.

P11. Graph Data Augmentation

- **Problem:** Graph data augmentations have been demonstrated to be beneficial for various graph learning tasks, but current graph data augmentations usually ignore the structure of the graph. Instead of doing random edge dropping, we can investigate a structure-aware distance for the data augmentations on the graphs. This problem is basically two-fold (1) how to define a structure-aware distance on the graph (2) how to sample/generate augmentations using this distance.
- **Data:** Many datasets possible, recommendation for the first paper below
- **Introductory papers:**
  - You, Yuning, et al. Graph Contrastive Learning with Augmentations. NeurIPS, 2020. arXiv preprint arXiv:2010.13902.
  - Zhao, Tong, et al. Data Augmentations for Graph Neural Networks. AAAI, 2021. arXiv preprint arXiv:2006.06830.
  - Suresh, Susheel, et al. Adversarial Graph Augmentation to Improve Graph Contrastive Learning. NeurIPS, 2021.
- **Comments:** Consider either parametric or non-parametric (e.g., graph optimal transport) distances.