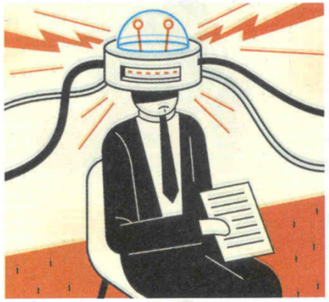
Programming language

- 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.

P1: Learning from fMRI Brain Imaging

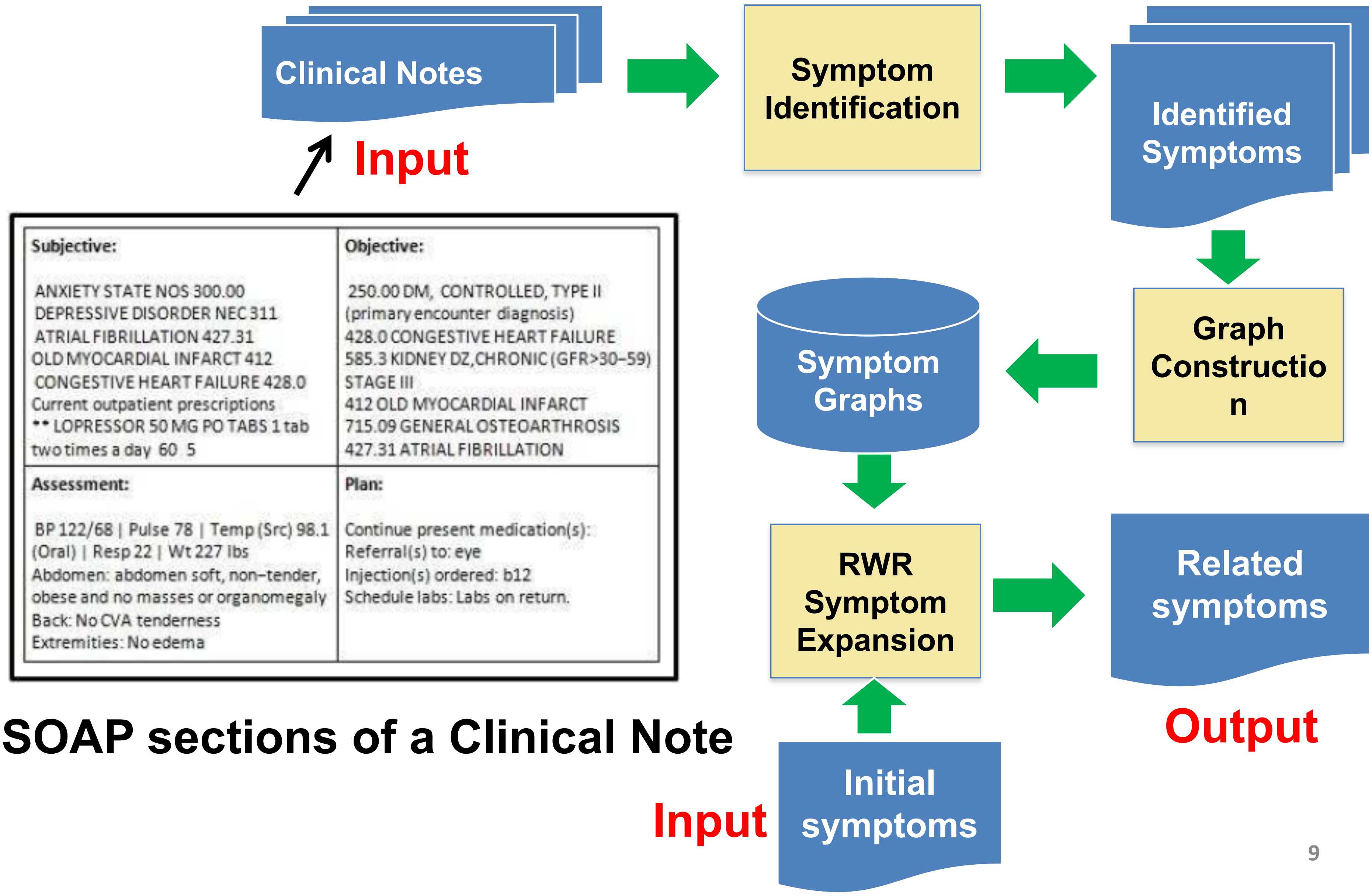


- ***Problem***: (1) Prediction Task: to predict when the subject was reading a sentence versus perceiving a picture. (2) Feature selection/reduction/invention, how to select most predictive features and/or how to come up new features to improve the prediction accuracy.
- ***Data***: StarPlus data. (http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-81/www/) This data set contains a time series of images of brain activation, measured using fMRI, with one image every 500 msec. During this time, human subjects performed 40 trials of a sentence-picture comparison task (reading a sentence, observing a picture, and determining whether the sentence correctly described the picture). Each of the 40 trials lasts approximately 30 seconds. Each image contains approximately 5,000 voxels (3D pixels), across a large portion of the brain. Data is available for 12 different human subjects. ADNI Data: http://www.adni-info.org/
- ***Introductory paper(s)***:
  - Mitchell et al, 2004: http://www.cs.cmu.edu/~tom/mlj04-final-published.pdf
  - Jiayu Zhou, Jun Liu, Vaibhav A. Narayan, Jieping Ye: Modeling disease progression via fused sparse group lasso. KDD 2012: 1095-1103

P2: Learning from Healthcare Web Sites

- ***Problem***: Find (extract and/or extend) relevant symptoms for a given disease
- ***Data***: Many Healthcare related web sites, e.g., Patient Like me, Diabetes Forum Data, WebMD.
- ***Introductory paper(s)***:
  - Parikshit Sondhi, Jimeng Sun, Hanghang Tong, ChengXiang Zhai: SympGraph: a framework for mining clinical notes through symptom relation graphs. KDD 2012: 1167-1175
- ***Comments***. High impact. Might lead to publications and/or headlines in news.

P2: Learning from Clinical Notes



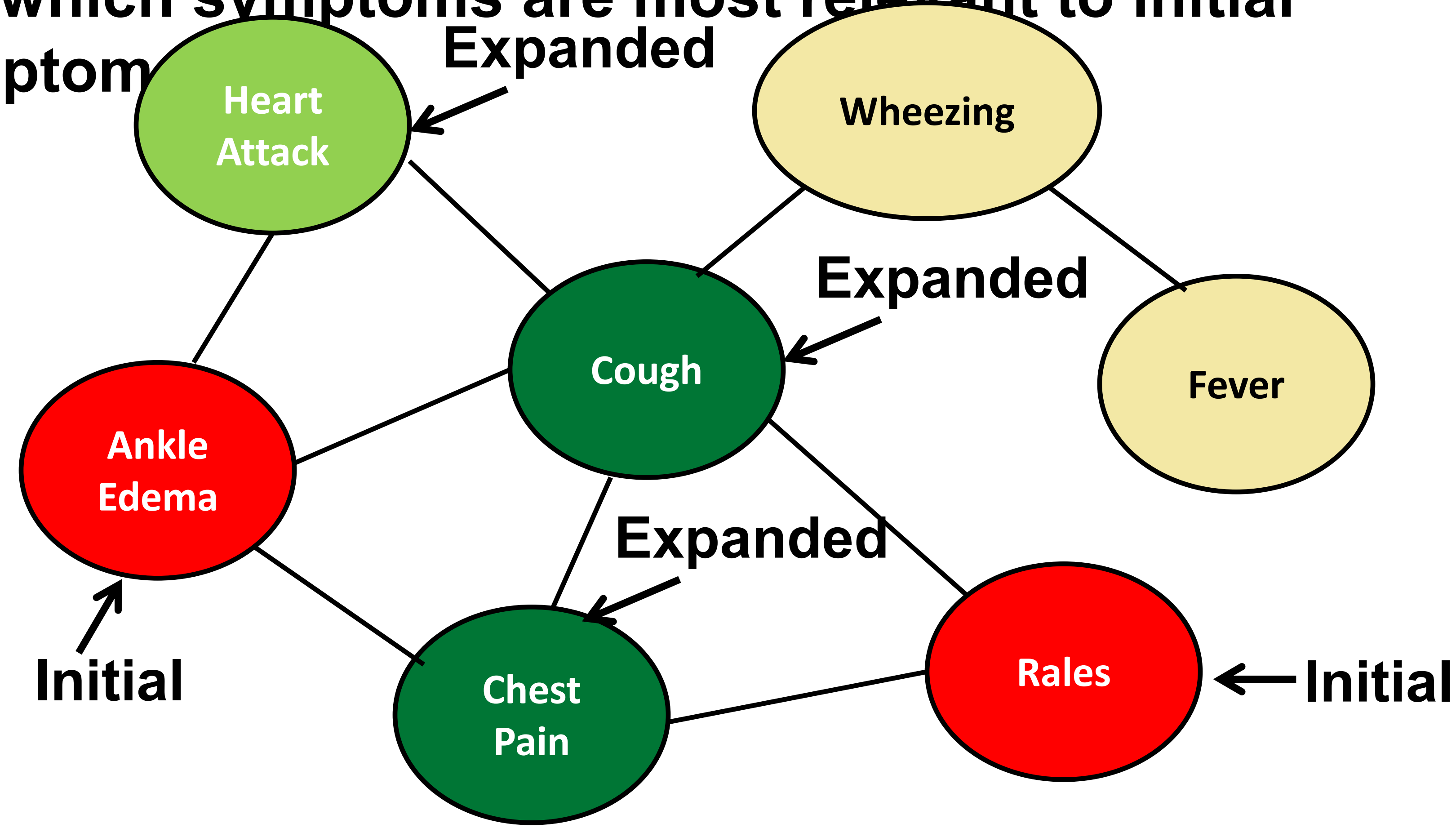
Subjective:	Objective:
ANXIETY STATE NOS 300.00 DEPRESSIVE DISORDER NEC 311 ATRIAL FIBRILLATION 427.31 OLD MYOCARDIAL INFARCT 412 CONGESTIVE HEART FAILURE 428.0 Current outpatient prescriptions ** LOPRESSOR 50 MG PO TABS 1 tab two times a day 60 5	250.00 DM, CONTROLLED, TYPE II (primary encounter diagnosis) 428.0 CONGESTIVE HEART FAILURE 585.3 KIDNEY DZ,CHRONIC (GFR>30-59) STAGE III 412 OLD MYOCARDIAL INFARCT 715.09 GENERAL OSTEOARTHROSIS 427.31 ATRIAL FIBRILLATION
Assessment:	**Plan:**
BP 122/68 | Pulse 78 | Temp (Src) 98.1 (Oral) | Resp 22 | Wt 227 lbs Abdomen: abdomen soft, non-tender, obese and no masses or organomegaly Back: No CVA tenderness Extremities: No edema	Continue present medication(s): Referral(s) to: eye Injection(s) ordered: b12 Schedule labs: Labs on return.

SOAP sections of a Clinical Note

P2: Learning from Clinical Notes: Symptom Expansion

Key Idea: Symptom Expansion → graph node proximity.

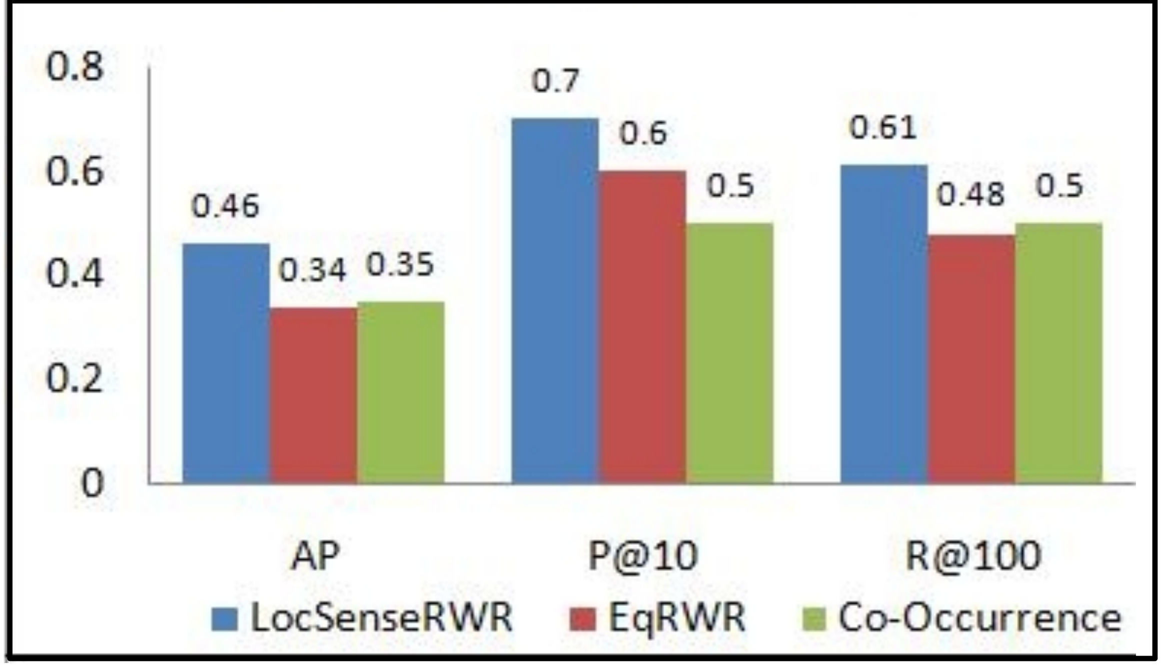
i.e., which symptoms are most relevant to initial symptom



Graph Proximity Algorithms

P2: Evaluations

Framingham Symptom Expansion



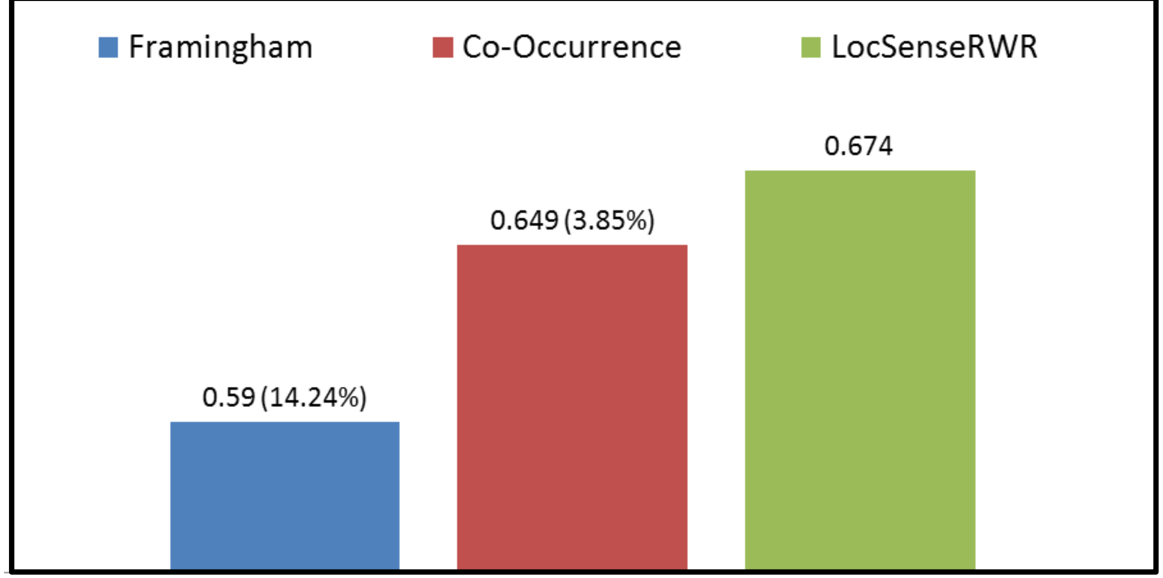
Evaluation Details:

Experts: 2; 175 symptoms judged

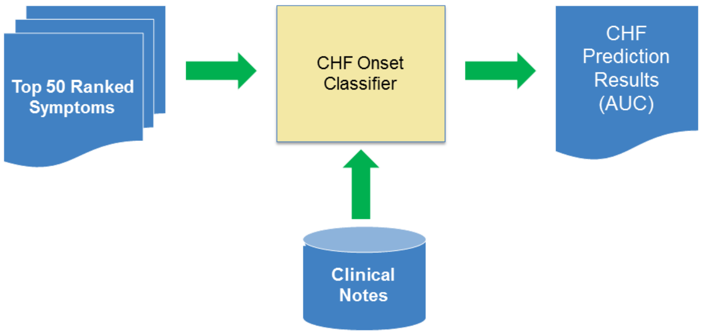
Relevant: 72 , Irrelevant: 103
Inter-annotator agreement : 81.8%

Symptoms labeled as related by both experts were considered as relevant.

CHF Prediction (AUC)

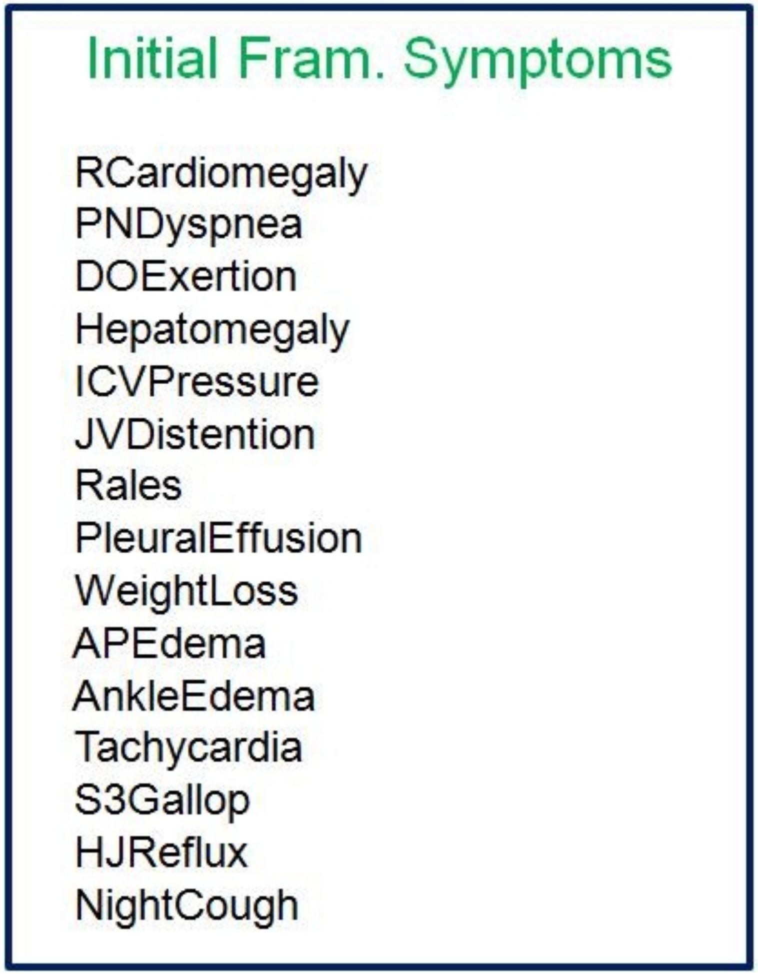


Evaluation Details:



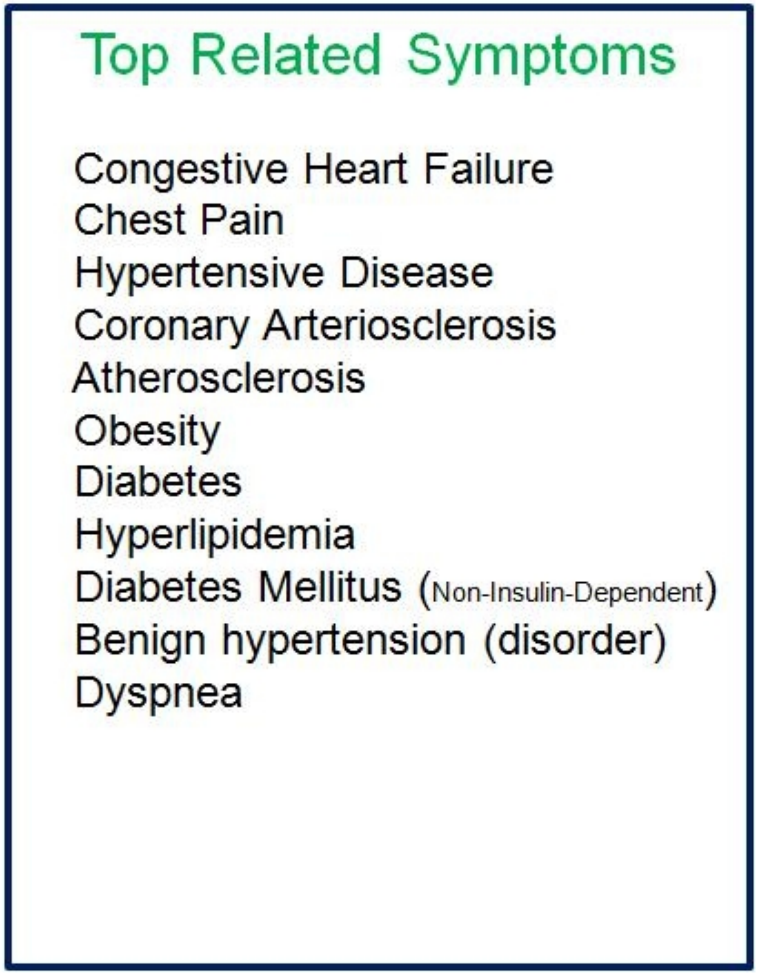
CHF: affecting 1 out of 5 adults in US; most costly in CMS
Framingham, 1971 → 50s, 60s

P2: Framingham Symptom Expansion



Initial Fram. Symptoms

RCardiomegaly
PNDyspnea
DOExertion
Hepatomegaly
ICVPressure
JVDistention
Rales
PleuralEffusion
WeightLoss
APEdema
AnkleEdema
Tachycardia
S3Gallop
HJReflux
NightCough

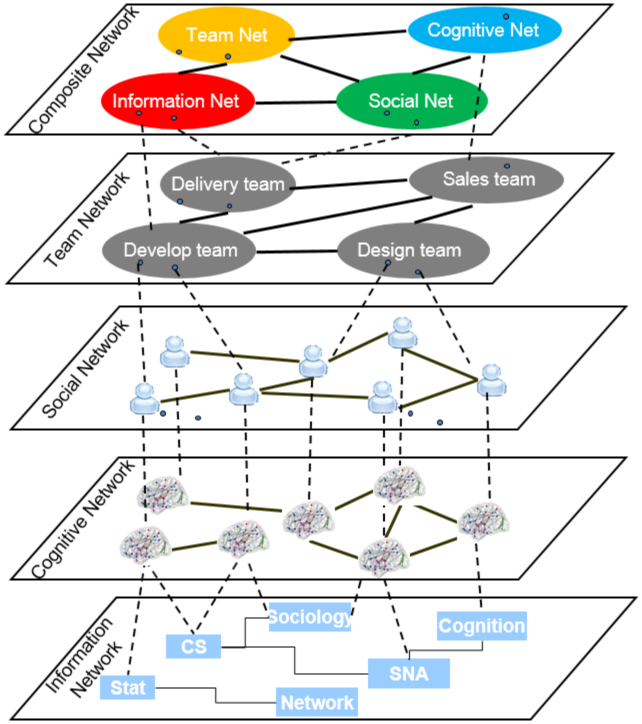


Top Related Symptoms

Congestive Heart Failure
Chest Pain
Hypertensive Disease
Coronary Arteriosclerosis
Atherosclerosis
Obesity
Diabetes
Hyperlipidemia
Diabetes Mellitus (Non-Insulin-Dependent)
Benign hypertension (disorder)
Dyspnea

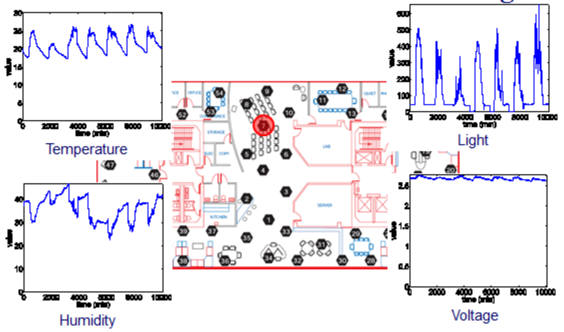
We are able to identify related disease and symptom mentions, confirmed by clinical experts

P3: Learning from Multiple Network



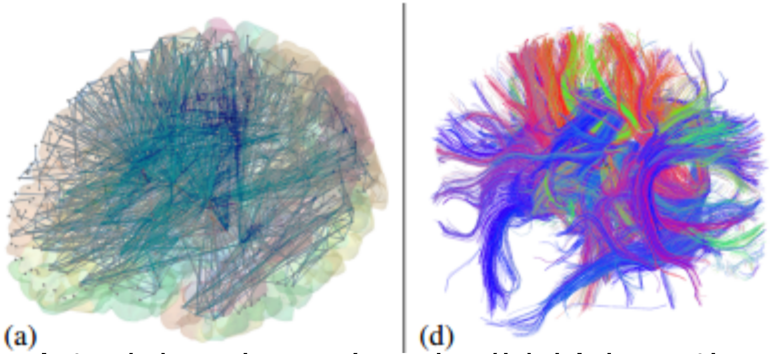
- ***Problem***: Given a set of networks (e.g., social network from facebook, twitter social networks, etc), how to model them and find interesting patterns.
- ***Data***: Refer to P1
- ***Introductory paper(s)***:
  - – Jingchao Ni, Hanghang Tong, Wei Fan, Xiang Zhang: Inside the atoms: ranking on a network of networks. KDD 2014: 1356-1365
- ***Comments***: Very hot topics in web/network science in the recent years. Many possible extension. Well likely to lead to publications.

P4: Learning from Sensor Data



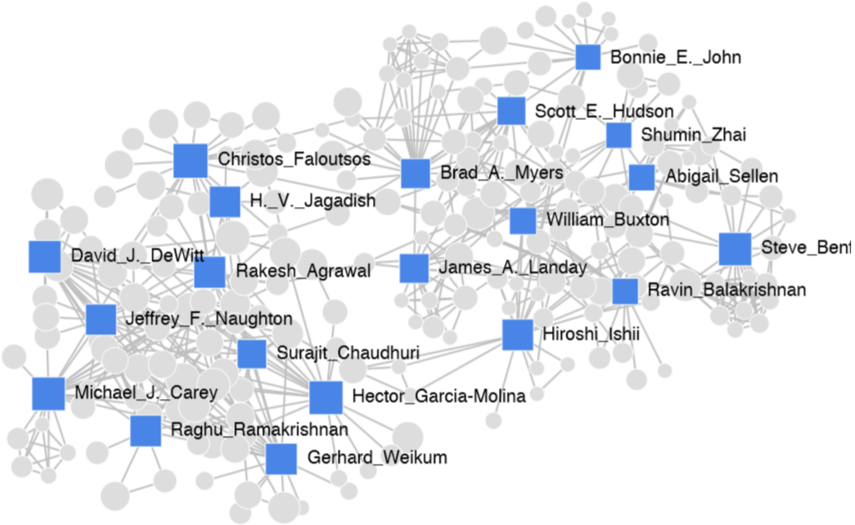
- ***Problem***: Given a set of multiple inter-correlatd time series data, how to find anomalies, patterns, etc; predict the future trend, impute missing values.
- ***Data***:(http://www.cs.cmu.edu/~guestrin/Research/Data/) This dataset contains temperature, humidity, and light data measurements, along with the voltage level of the batteries at each node, using this 54-node sensor network deployment. The data was collected every 30 seconds, starting around 1am on February 28th 2004.
- ***Introductory paper(s)***:
  - Cai et al. 2015: http://ycai.ws.gc.cuny.edu/files/2015/03/NoT_sdm15.pdf
  - Guestrin et al. 2004: http://www.cs.cmu.edu/~guestrin/Publications/IPSN2004/ipsn2004.pdf
  - Deshpande et al. 2004: http://www.cs.cmu.edu/~guestrin/Publications/VLDB04/vldb04.pdf
- ***Comments***: This is a "real" dataset, with lots of missing data, noise, and failed sensors giving outlier values, especially when battery levels are low.

P5: Brain Network Analysis

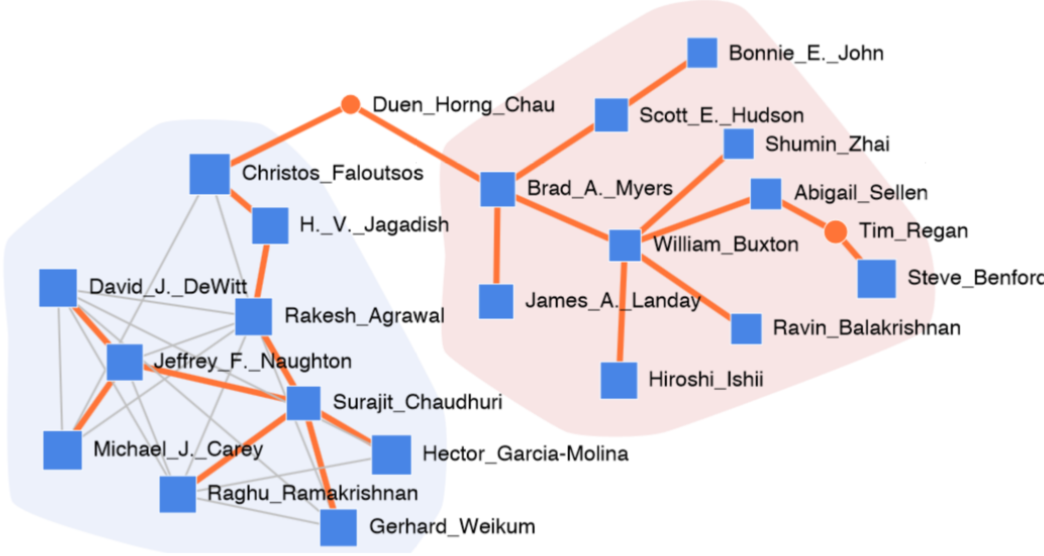


- ***Problem***: (1) Prediction Task: Classify human-subjects into (a) male vs. female, (b) high-math capable vs. normal, (c) creative vs. normal. (2) again, feature selection/invention, etc. (3) Visual analytics: how to visualize and compare two or more brain networks (e.g., what is the commonality, what is the difference, etc).
- ***Data***: brain network data: (https://www.andrew.cmu.edu/user/lakoglu/courses/95828/S17/projectsources/brainnetworks.rar)
  This dataset contains the brain connectivity graphs of 114 human subjects. Each brain is segmented into 70 regions (or super-voxels). The network depicts the connectivity among these regions, where weights on links represent the strength of the connection. Meta-data on human-subjects include gender, age, IQ, etc. as well as scores obtained by tests evaluating the math capability or creativity of the subjects.
- ***Introductory paper(s)***:
  - Duarte-Carvajalino, 2012:
    http://www.sciencedirect.com/science/article/pii/S1053811911012687
  - Alper et al 2013]:
    http://research.microsoft.com/en-us/um/people/nath/docs/brainvis_chi2013.pdf
- ***Comments***. Very related to BRAIN Initiative, read this (https://www.whitehouse.gov/share/brain-initiative). Might lead to publications and/or headlines in news.

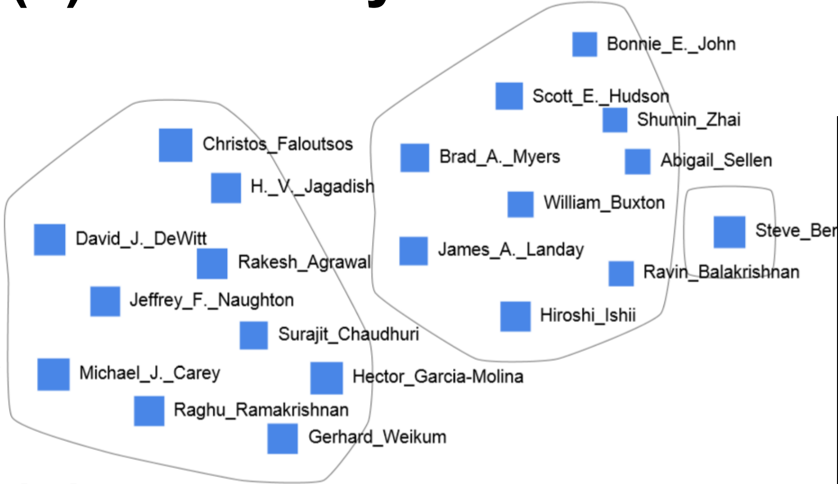
P6. From Pattern Match to Sense-Making



(a) Too many connections?



(c) Our sol.: ‘right’ connections → better sense-making



(b) Too few connections?

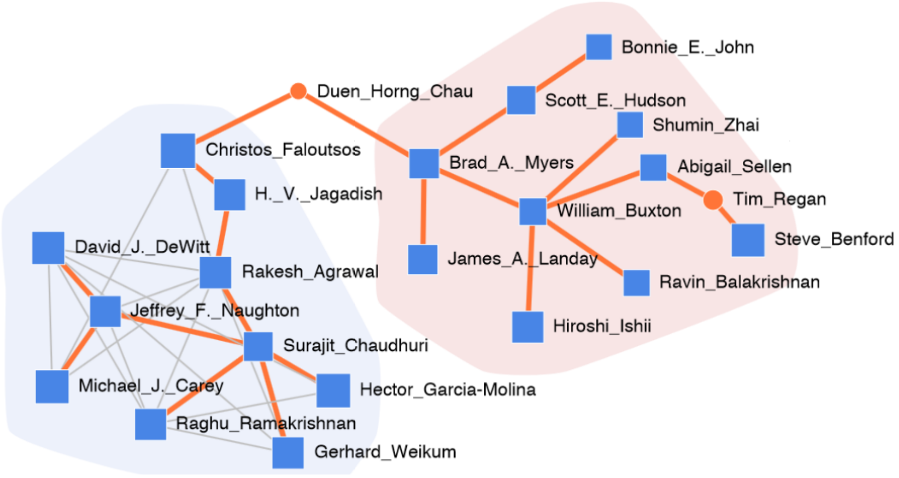
+ ‘right’ connections = most succinct way to describe marked nodes

+ MDL-based formulation

+ NP-Hard (Reduction from Steiner tree)

P6. From Pattern Match to Sense-Making

Problem:



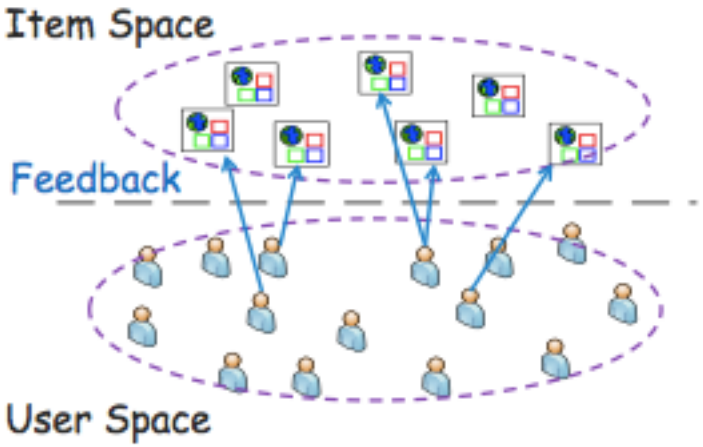
***Data*: DBLP database (http://dblp.uni-trier.de/xml/); IMDB datasets (http://www.infochimps.com/tags/imdb); the ENRON dataset (https://www.cs.cmu.edu/~enron/)**

***Introductory paper(s)*:**

- Leman Akoglu, Duen Horng Chau, Christos Faloutsos, Nikolaj Tatti, Hanghang Tong, Jilles Vreeken: Mining Connection Pathways for Marked Nodes in Large Graphs. SDM 2013: 37-45
- Duen Horng Chau, Leman Akoglu, Jilles Vreeken, Hanghang Tong, Christos Faloutsos: TourViz: interactive visualization of connection pathways in large graphs. KDD 2012: 1516-1519

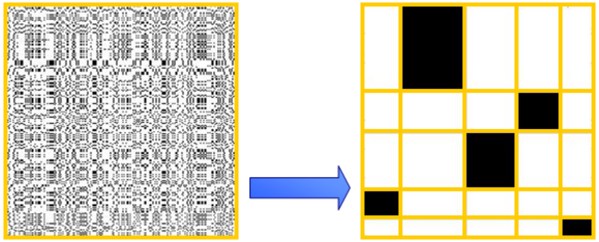
***Comments*:** Extremely useful tools in many application domains. Well likely to lead to the real systems. Might lead to publications.

P7. Recommendation



- ***Problems:*** (think of Netflix competition)
- ***Introductory Paper(s):***
  - Yehuda Koren, Robert M. Bell, Chris Volinsky: Matrix Factorization Techniques for Recommender Systems. IEEE Computer 42(8): 30-37 (2009)
  - Yuan Yao, Hanghang Tong, Guo Yan, Feng Xu, Xiang Zhang, Boleslaw K. Szymanski, Jian Lu: Dual-Regularized One-Class Collaborative Filtering. CIKM 2014: 759-768
- ***Data***: Netflix competition as well as those used in the above paper
- ***Comments***: very very hot topics. Many possible directions (one class; side information, code-start, dynamics, social recommendation, attribution).

P8: Cross-Association/Co-clustering



- ***Problem***: Given a contingency table (i.e. large sparse bipartite graph w/ positive edge weights), how can we simultaneously group the row/column into some clustering? How do we do that in a parameter-free way? Recently, researchers try to solve this problem from information theoretic point of view, by exploring the duality between data mining and compression. One challenge is how to generalize it to heterogeneous settings, for example, when we have 3 types of nodes (in DBLP: authors, papers, keywords), and we want to find groups and clusters.
- ***Data:*** DBLP database (http://dblp.uni-trier.de/xml/); IMDB datasets (http://www.infochimps.com/tags/imdb); the ENRON dataset (https://www.cs.cmu.edu/~enron/)
- ***Introductory paper(s)***:
  - Inderjit S. Dhillon: Co-clustering documents and words using bipartite spectral graph partitioning. KDD 2003
  - Deepayan Chakrabarti, Spiros Papadimitriou, Dharmendra S. Modha, Christos Faloutsos: Fully automatic cross-associations. KDD 2004: 79-88
  - Bo Long, Zhongfei (Mark) Zhang, Philip S. Yu: Combining Multiple Clusterings by Soft Correspondence. ICDM 2005: 282-289
- ***Comments***: There are several possible extensions of the current methods. Might lead to publications. There is also some practical issues, for example, efficient implementation of the current methods, in a conventional machine, or on a 'hadoop' cluster, see this paper for an example.

P9: Graph Learning -Information Dissemination

- ***Problem:*** We want to find patters of propagation of information (or viruses, influence, etc.) in a network. For example, in a web-log influence tree, *what is the most typical form of influence*: a 'star' topology? a 'string' topology? something in-between? How to generate such realistic patterns, from first principles? Also, we want to model the temporal aspects: how often do bloggers post messages? *are the posts uniformly distributed over time?* (probably not, probably bursty). How can we spot abnormal/surprising patterns? How can we affect the outcome of dissemination?
- ***Data:*** social networks, citation networks, weblog influence data.
- ***Introductory paper(s)***:
  - Jure Leskovec, Mary McGlohon, Christos Faloutsos, Natalie S. Glance, Matthew Hurst: Patterns of Cascading Behavior in Large Blog Graphs. SDM 2007: 551-556
  - Hanghang Tong, B. Aditya Prakash, Charalampos E. Tsourakakis, Tina Eliassi-Rad, Christos Faloutsos, Duen Horng Chau: On the Vulnerability of Large Graphs. ICDM 2010: 1091-1096
  - Hanghang Tong, B. Aditya Prakash, Tina Eliassi-Rad, Michalis Faloutsos, Christos Faloutsos: Gelling, and melting, large graphs by edge manipulation. CIKM 2012: 245-254

- ***Comments***: high-impact work, might lead to publication.

P10: Team Performance and Formation

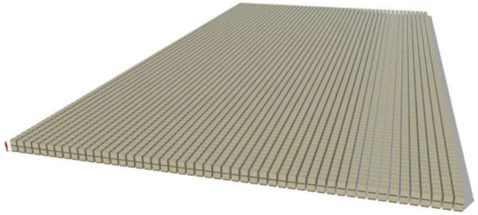
- ***Problem:*** (1) what makes a successful team? (2) how to predict the performance of a give team (e.g., the outcome of a game); (3) how to form a good team?
- ***Data***
  - NBA statistics (http://www.cs.stonybrook.edu/~leman/courses/13CSE512/projectsources/databasebasketball2.0.zip)
  - citation data (http://arnetminer.org/citation)
  - IMDB data, Scholarly data.
- ***Introductory paper(s)***:
  - Theodoros Lappas, Kun Liu, Evimaria Terzi: Finding a team of experts in social networks. KDD 2009: 467-476
  - Liangyue Li, Hanghang Tong, Nan Cao, Kate Ehrlich, Yu-Ru Lin and Norbou Buchler: Replacing the Irreplaceable: Fast Algorithms for Team Member Recommendation. WWW 2015
- ***Comments***: high-impact work, might lead to publication.

P11: Learning from CQA data

- ***Problem**:* (1) question routing – how to route a given question to the 'right' expert? (2) question-answer tagging – how to organize the question-answer pair in a more structured way? (3) how to spot a good/bad question/answer? (4) what makes a user to stay/leave the forum?
- ***Data**:* Stackoverflow data (http://blog.stackoverflow.com/category/cc-wiki-dump/).

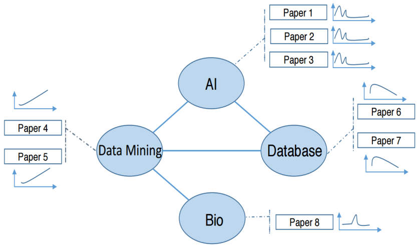
  The content/text of questions/answers, the voting (upvote/downvote), etc.
- ***Introductory paper(s)***:
  - Ashton Anderson, Daniel P. Huttenlocher, Jon M. Kleinberg, Jure Leskovec: **Discovering value from community activity on focused question answering sites: a case study of stack overflow.** KDD 2012: 850-858
  - Li, Wei, Charles Zhang, and Songlin Hu. "G-Finder: routing programming questions closer to the experts." *ACM Sigplan Notices* 45.10 (2010): 62-73.
  - Yuan Yao, Hanghang Tong, Feng Xu, Jian Lu: Predicting long-term impact of CQA posts: a comprehensive viewpoint. KDD 2014: 1496-1505
  - Richter, Yossi, Elad Yom-Tov, and Noam Slonim. "Predicting customer churn in mobile networks through analysis of social groups." *Proceedings of the 2010 SIAM international conference on data mining*. Society for Industrial and Applied Mathematics, 2010.
  - Jagat Sastry Pudipeddi, Leman Akoglu, Hanghang Tong: User churn in focused question answering sites: characterizations and prediction. WWW 2014: 469-474
- ***Comments***: very rich data set, many possible directions, might lead to publication.

P12: Anomaly Detection



- ***Problem**:* how to find rare/unusual instances/groups?

- ***Data**:* Many possible data sets, see the data sets in the previous data sets as well as those in papers below.
- ***Introductory paper(s)***:
  - Shashank Pandit, Duen Horng Chau, Samuel Wang, Christos Faloutsos: NetProbe: A Fast and Scalable System for Fraud Detection in Online Auction Networks. WWW 2007
  - Leman Akoglu, Hanghang Tong, Jilles Vreeken, Christos Faloutsos: Fast and reliable anomaly detection in categorical data. CIKM 2012: 415-424
  - Jingrui He, Hanghang Tong, Jaime G. Carbonell: Rare Category Characterization. ICDM 2010: 226-235
- ***Comments***: open-ended, very broad applicability, might lead to publication.



P13: Scholarly Impact Forecasting

- ***Problem:*** Given the bibliographic information of scholarly publications, predict the long-term impact or impact pathway for the scholarly entities (papers, authors, venues). The impact could be citation counts, *h*-index, impact factor and many more.

- ***Data**:* Aminer citation network (https://aminer.org/citation)

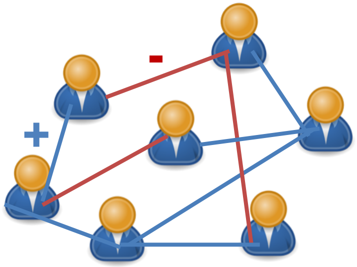
- ***Introductory paper(s)***:
  - Liangyue Li, Hanghang Tong. The Child is Father of the Man: Foresee the Success at the Early Stage. KDD, 2015.
  - Liangyue Li, Hanghang Tong, Jie Tang, Wei Fan. iPath: Forecasting the Pathway to Impact. SDM, 2016.

- ***Comments:*** many facets could be considered into the model, e.g., non-linearity, correlation, temporal smoothness, etc.

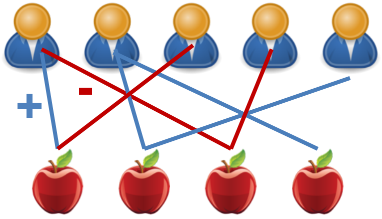
P14: Learning from Signed Networks

- ***Problem***: In real applications, links in the networks may not imply positive connections between entities. On the contrary, some of them may indicate negative relationships between nodes. For example, in social networks, friends are connected by positive links while enemies are marked by negative links. While in recommender systems, positive reviews from the users to the products are positive links while negative reviews are represented by negative links. Utilized the sign network can improve the performance of various tasks like link prediction, product recommendation, etc. Moreover, inferring an accurate signed network is also an important task in related literature.
- ***Data***:
  - Epinion dataset: http://www.cse.msu.edu/~tangjili/trust.html
  - Slashdot dataset: https://snap.stanford.edu/data/
- ***Introductory paper(s)***:
  - Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. 2010a. Predicting positive and negative links in online social networks. In Proceedings of the 19th international conference on World wide web. ACM, 641–650
  - Jiliang Tang, Charu Aggarwal, and Huan Liu. 2016b. Recommendations in signed social networks. In Proceedings of the 25th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 31–40.
- ***Comments***: still has many open problems to solve, can be applied to many applications, may lead to publications

Social Networks:



Recommender Systems:



— : Positive Links

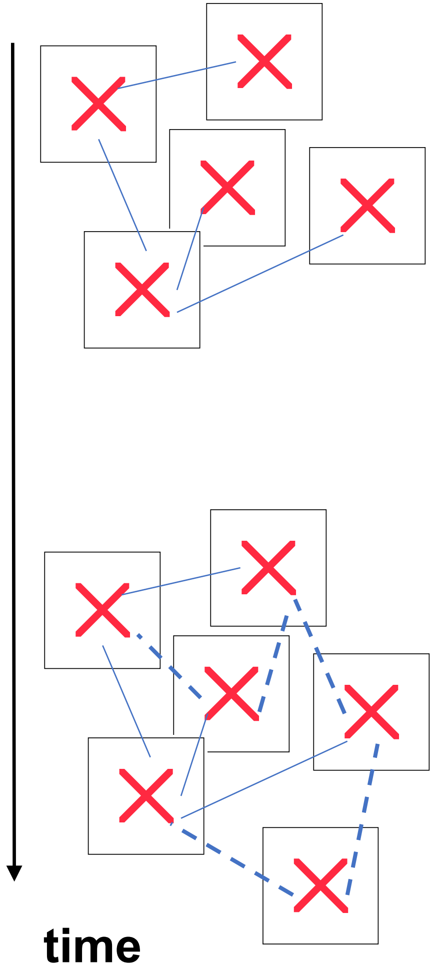
— : Negative Links

P15: Multi-label classification for graph nodes

- **Problem:** In this problem, we want to assign one or more labels from a finite label set to each node in the graph. We have observed a certain fraction of nodes and all their labels, and the task is to predict the labels for the remaining nodes in the graph. In this project, possible challenges are: (1) how can we use the structural information of the nodes, and (2) how can we leverage the attribute information of the nodes?
- **Introductory papers:**
  - Grover, Aditya, and Jure Leskovec. "node2vec: Scalable Feature Learning for Networks.“
  - Perozzi, Bryan, Rami Al-Rfou, and Steven Skiena. "Deepwalk: Online learning of social representations." *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014.
- **Data:** social networks, PPI networks (e.g., datasets used in the above papers)
- **Comments:** very hot topics in recent years, might lead to publications.

P16: Learning from dynamic networks

- **Problem:** Real-world networks are often evolving over time, with the addition/deletion of nodes and edges (e.g., user friendships, coauthor relations), and the changes of associated node/edge attributes (user profile, research interests). We are interested in developing novel and principled algorithms to gain insights from the dynamic networks in near real-time. Some interesting research problems can be how to find node communities structure changes, anomalous nodes, predicting the formation of new links
- **Datasets:** DBLP, Aminer, Twitter, Facebook
  - Aminer: https://aminer.org/data
  - Streamspot: http://www3.cs.stonybrook.edu/~emanzoor/streamspot/
- **Related papers:**
  - Manzoor, Emaad, Sadegh M. Milajerdi, and Leman Akoglu. "Fast memory-efficient anomaly detection in streaming heterogeneous graphs." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016.
  - Tang, Lei, et al. "Community evolution in dynamic multi-mode networks." *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008.
  - Zhao, Peixiang, Charu Aggarwal, and Gewen He. "Link prediction in graph streams." *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*. IEEE, 2016.



P17: Fairness in Machine Learning

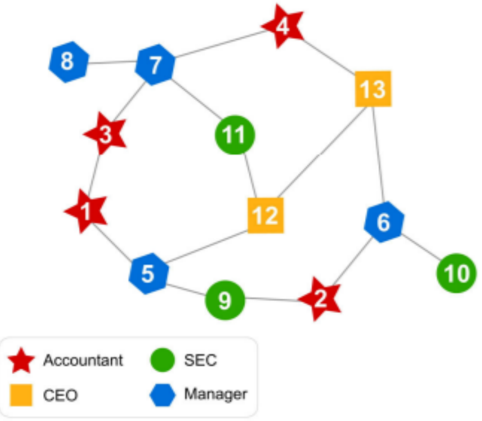
- **Problem:** Automated decision-making process is increasing the risk of discrimination against certain groups of people (e.g. race, gender). Researchers try to find methods to detect algorithmic transparency and to debias data and model. Can you develop your own method to (1) detect unfairness and (2) make fair decisions without discrimination?
- **Data:** adult income dataset (https://archive.ics.uci.edu/ml/machine-learning-databases/adult/)

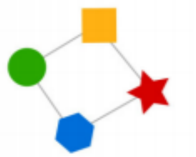
  German credit data (https://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german/)

  criminal history dataset (https://github.com/propublica/compas-analysis/)
- **Introductory resources**
  - Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. KDD 2015.
  - Datta A., Sen S., Zick Y. (2017) Algorithmic Transparency via Quantitative Input Influence. In: Cerquitelli T., Quercia D., Pasquale F. (eds) Transparent Data Mining for Big and Small Data. Studies in Big Data, vol 32. Springer, Cham
  - Fairness in Machine Learning [https://fairmlclass.github.io/]
- **Comments:** High impact research. Very hot topics in recent years. Well likely to lead to publication.

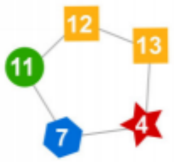
P18 Learning interesting patterns from network

- **Problem**: In many real applications, we need to find matching subgraphs as an answer for some query patterns/templates from a larger network. The matching subgraphs could be exact (isomorphic) to the query or not necessarily exact but similar to some extent. The query patterns could contain special properties or arbitrary. How to use machine learning techniques to detect such patterns is an important task in ML/DM field.
- **Data**: Stanford large network data collection (http://snap.stanford.edu/data/index.html), datasets in the introductory papers.
- **Introductory papers:**
  - [1] Tong, Hanghang, et al. "Fast best-effort pattern matching in large attributed graphs." *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2007.
  - [2] Du, Boxin, et al. "First: Fast interactive attributed subgraph matching." *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2017.
- **Comment**: Very important application problem in data mining/machine learning area, may lead to publication.





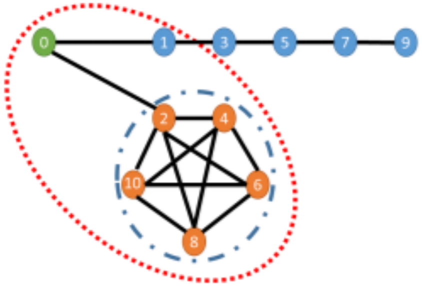
(a) Loop query



(b) A matching subgraph

P19: Mining rich sub-structures in network

- **Problem**: In many data mining/machine learning applications, people are interesting in finding subgraphs which contain large amount of certain structures (such as triangles, loops) from a larger network. This problem is often modeled as a clustering or graph cut problem. But different from traditional clustering/graph cut problem, the clusters/graph cuts should meet certain requirements to be able to contain those sub-structures, so the traditional methods should be altered accordingly.
- **Data**: datasets in the introductory papers.
- **Introductory papers:**
  - [1] D Zhou, S Zhang, M.Y. Yildirim, S Alcorn, H Tong, H Davulcu, J He. A Local Algorithm for StructurePreserving Graph Cut, ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2017
  - [2] Yin, Hao, et al. "Local higher-order graph clustering." *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2017.
- **Comment**: Very hot topic in recent years, may lead to publication



A cluster that contains rich triangles

P20: Android malware detection through structured information network

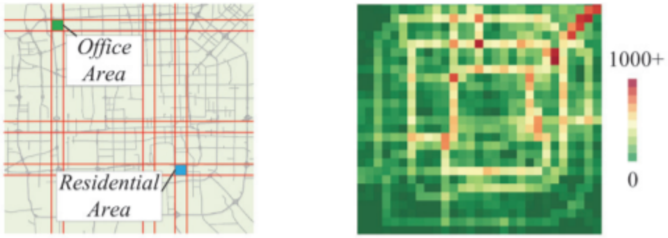
- Problem: With explosive growth of Android malware and due to the severity of its damages to smart phone users, the detection of Android malware has become an increasingly important topic in cyber security. One of the specific problem is to detect Android malware based on the heterogeneous structured information network, instead of just using Application Programming Interface (API) calls only. In this project you are supposed to use a meta-path based approach to characterize the semantic relatedness of apps and APIs (please refer to [1]), and come up with your own method to detect malwares based on this relationship.
- Data: datasets in the introductory papers.
- Introductory papers:
  - [1] Hou, Shifu, et al. "Hindroid: An intelligent android malware detection system based on structured heterogeneous information network." *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2017.
- Comment: very practical and hot topic in the data mining/machine learning area, may lead to publication

P21: Part-Whole outcome prediction

- ***Background****:* part-whole relationship routinely finds itself in many application scenarios, e.g.,
  - *Teams*: whole is team, parts are individual team members
  - *CQA*: whole is question, parts are individual answers the question receives
  - *Stock*: whole is market index, parts are individual stocks
- ***Problem***: how to predict the outcome of both the whole and parts (e.g., team's performance/team member's performance)
- ***Data****:* Aminer citation network (https://aminer.org/citation)

  Stack Overflow (http://data.stackexchange.com)
- ***Introductory paper(s)***:
  - Liangyue Li, Hanghang Tong, Yong Wang, Conglei Shi, Nan Cao and Norbou Buchler. Is the Whole Greater Than the Sum of Its Parts? KDD, 2017.
  - Yuan Yao, Hanghang Tong, Feng Xu, Jian Lu. Predicting long-term impact of CQA posts: a comprehensive viewpoint. KDD, 2014.
- ***Comments:*** the outcome of the parts and whole may not be independent of each other

P22: Traffic flow prediction



(a) Grid-based map segmentation (b) Inflow matrix

- **Problem**: In this problem, we are given an inflow matrix and an outflow matrix, which measure the traffic flow on the grid-based map over time. We want to predict the flow of the map in the future. Possible challenges are: (1) how to use both the spatial and temporal dependencies to predict the flow? (2) how to deal with the external factors which may shortly affect the flow, e.g., weather conditions, special events, etc.
- **Introductory papers:**
  - Lv, Yisheng, et al. "Traffic flow prediction with big data: A deep learning approach." *IEEE Trans. Intelligent Transportation Systems* 16.2 (2015): 865-873.
  - Zhang, Junbo, Yu Zheng, and Dekang Qi. "Deep Spatio-Temporal Residual Networks for Citywide Crowd Flows Prediction." *AAAI*. 2017.
- **Datasets**:
  - datasets in the above papers (e.g.,https://github.com/lucktroy/DeepST/tree/master/data/TaxiBJ)
- **Comments**: very hot topics, might lead to publications