IP1

Tba - Goel

Abstract not available at time of publication.

Ashish Goel
Stanford University
ashishg@stanford.edu

IP2

Tba - Jones

Abstract not available at time of publication.

Steve Jones
BC Cancer Agency and Simon Fraser University
University of British Columbia
sjones@bcgsc.ca

IP3

Tba - Chaudhuri

Abstract not available at time of publication.

Surajit Chaudhuri
Microsoft Research
surajitc@microsoft.com

IP4

Tba - Yu

Abstract not available at time of publication.

Bin Yu
University of California, Berkeley
binyu@stat.berkeley.edu

CP1

Where Graph Topology Matters: The Robust Subgraph Problem

Robustness is a critical measure of the resilience of large networked systems. Most prior works focus on the global robustness of a given graph at large, e.g., by measuring its overall vulnerability to attacks or failures. In this paper, we pose a novel problem: given a large graph, how can we find its most robust local subgraph (RLS)? Our formulation is related to the general framework [39] for the densest subgraph problem, however differs from it substantially, as robustness also concerns with the placement of edges, i.e., topology. We show that the RLS-PROBLEM is NP-hard and propose two heuristic algorithms based on top-down and bottom-up search strategies. Experiments demonstrate that we find subgraphs with larger robustness than the densest subgraphs [8, 39] even at lower densities, suggesting that the existing approaches are not suitable for the new problem setting.

Leman Akoglu
Stonybrook University
leman@cs.stonybrook.edu

Hau Chan, Shuchu Han
Stony Brook University
hauchan@cs.stonybrook.edu, shhan@cs.stonybrook.edu

CP1

Selecting Shortcuts for a Smaller World

The small world phenomenon is a desirable property of social networks, since it guarantees short paths between the nodes of the social graph and thus efficient information spread on the network. Thus, small world property benefits both network users and network owners. In this work, we study the problem of finding a subset of k edges from a set of candidate edges whose addition to a network leads to the greatest reduction in its average shortest path length.

Nikos Parotsidis, Evaggelia Pitoura
University of Ioannina, Greece
nparotsi@cs.uoi.gr, pitoura@cs.uoi.gr

Panayiotis Tsaparas
University of Ioannina
tsap@cs.uoi.gr

CP1

Significant Subgraph Mining with Multiple Testing Correction

The problem of finding itemsets that are statistically significantly enriched in a class of transactions is complicated by the need to correct for multiple hypothesis testing. Pruning *untestable hypotheses* was recently proposed as a strategy for this task of significant itemset mining. It was shown to lead to greater statistical power, the discovery of more truly significant itemsets, than the standard Bonferroni correction on real-world datasets. An open question, however, is whether this strategy of excluding untestable hypotheses also leads to greater statistical power in *subgraph* mining, in which the number of hypotheses is much larger than in itemset mining. Here we answer this question by an empirical investigation on eight popular graph benchmark datasets. We propose a new efficient search strategy, which always returns the same solution as the state-of-the-art approach and is approximately two orders of magnitude faster. Moreover, we exploit the dependence between subgraphs by considering the *effective number of tests* and thereby further increase the statistical power.

Mahito Sugiyama
Max Planck Institutes Tübingen
mahito@ar.sanken.osaka-u.ac.jp

Felipe Llinares Lopez
ETH Zurich
felipe.llinares@bsse.ethz.ch

Niklas Kasenburg
University of Copenhagen
niklas.kasenburg@di.ku.dk

Karsten Borgwardt
ETH Zurich
karsten.borgwardt@bsse.ethz.ch

CP1

Same Bang, Fewer Bucks: Efficient Discovery of the Cost-Influence Skyline

Influence maximization aims to find a set of persons in a social network that can be used as seeds for a viral market-

ing campaign, such that expected influence is maximized. Standard approaches to this problem produce a single seed set that either maximizes influence, or the "bang for the buck" if the vertices have costs. We consider the problem of finding the cost-influence skyline, i.e., the collection of all seed sets that are Pareto optimal w.r.t. seeding cost and expected influence. We show that computing the cost-influence skyline has advantages over finding a single solution only. Now, the problem is to discover the skyline w.r.t. two functions spanned by all subsets of size k of a set of vertices. This is a hard problem, hence we present an efficient heuristic algorithm for computing the skyline when one of the functions is linear and the other submodular. The experiments show that the cost-influence skyline can be computed for networks with up to a million vertices.

Antti Ukkonen
Yahoo! Research
antti.ukkonen@ttl.fi

Matthijs Van Leeuwen
KU Leuven
matthijs.vanleeuwen@cs.kuleuven.be

CP1

Functional Node Detection on Linked Data

Networks, which characterize object relationships, are ubiquitous in various domains. One very important problem is to detect the nodes of a specific function in these networks. For example, is a user normal or anomalous in an email network? Does a protein play a key role in a protein-protein interaction network? In this talk, a novel *F*eature *I*ntegration based Functional *N*ode *D*etection (FIND) algorithm is presented. Specifically, FIND extracts the most discriminative information from both node characteristics and network structures in the form of a unified latent feature representation with the guidance of several labeled nodes. Experiments on two real world data sets validate that the proposed method significantly outperforms the baselines on the detection of three different types of functional nodes.

Kang Li
The State University of New York at Buffalo
kli22@buffalo.edu

Jing Gao
University at Buffalo
jing@buffalo.edu

Suxin Guo
The State University of New York at Buffalo
suxinguo@buffalo.edu

Nan Du
State University of New York at Buffalo
nandu@buffalo.edu

Aidong Zhang
Department of Computer Science
State University of New York at Buffalo
azhang@buffalo.edu

CP2

Spectral Embedding of Signed Networks

Social networks are often modelled by graphs, with nodes representing individuals, and positively weighted edges representing the strength of the relationships between them. In a good embedding, edges that are heavily (positively) weighted, and so represent strong interactions, cause the vertices they connect to be embedded close to one another. However, in some social networks there are also antagonistic relationships that are naturally represented by negatively weighted edges. The resulting graphs are called signed graphs. Clearly an embedding of such a signed graph should place nodes connected by positively weighted edges close together, but nodes connected by negatively weighted edges far apart. Existing spectral techniques to embed signed graphs have serious drawbacks. We derive two normalized spectral analysis methods for signed graphs and show, using real-world data, that they produce robust embeddings.

David Skillicorn
Queen's University, Canada
skill@cs.queensu.ca

Quan Zheng
Queen's University
quan@cs.queensu.ca

CP2

Feature Selection for Nonlinear Regression and Its Application to Cancer Research

Feature selection is a fundamental problem in machine learning. With the advent of high-throughput technologies, it becomes increasingly important in a wide range of scientific disciplines. In this paper, we consider the problem of feature selection for high-dimensional nonlinear regression. This problem has not yet been well addressed in the community, and existing methods suffer from issues such as local minima, simplified model assumptions, high computational complexity and selected features not directly related to learning accuracy. We propose a new wrapper method that addresses some of these issues. We start by developing a new approach to estimating sample responses and prediction errors, and then deploy a feature weighting strategy to find a feature subspace where a prediction error function is minimized. We formulate it as an optimization problem within the SVM framework and solve it using an iterative approach. In each iteration, a gradient descent based approach is derived to efficiently find a solution. A large-scale simulation study is performed on four synthetic and nine cancer microarray datasets that demonstrates the effectiveness of the proposed method.

Yijun Sun
Department of Microbiology and Immunology
The State University of New York at Buffalo
yijunsun@buffalo.edu

CP2

Efficient Partial Order Preserving Unsupervised Feature Selection on Networks

In this paper, we investigate unsupervised feature selection problem on networks. To effectively incorporate linkage information, we propose a Partial Order Preserving (POP) principle for evaluating features. We show the advantage of this novel formulation in several respects: effectiveness, efficiency and its connection to optimizing AUC. We propose three instantiations derived from the POP principle and evaluate them using three real-world datasets. Experimental results show that our approach has significantly better

performance than state-of-the-art methods under several different metrics.

Xiaokai Wei, Sihong Xie
University of Illinois at Chicago
xwei2@uic.edu, sxie6@uic.edu

Philip S. Yu
University of Illinois at Chicago
Chicago, USA
psyu@uic.edu

CP2

From Categorical to Numerical: Multiple Transitive Distance Learning and Embedding

Categorical data are ubiquitous in real-world databases. However, due to the lack of an intrinsic proximity measure, many powerful algorithms for numerical data analysis may not work well on the categorical counterparts, making it a bottleneck in practical applications. In this paper, we propose a novel method to transform categorical data to numerical representations, in order to open the possibility of exploiting the abundant, numerical learning algorithms in a great variety of categorical data mining problems. Our key idea is to learn a pairwise dis-similarity among categorical symbols, henceforth a continuous embedding, which can then be used for subsequent numerical treatment. There are two important criteria for learning the dis-similarities. First, it should capture the important "transitivity" which has shown to be particularly useful in measuring the proximity relation in the categorical data. Second, the pairwise sample geometry arising from the learned symbol distances should be maximally consistent with prior knowledge (e.g., class labels) to obtain good generalization performance. We achieve these by designing a multiple transitive distance learning and embedding method. Encouraging results are observed on a number of benchmark classification tasks against state-of-the-art.

Kai Zhang
NEC Laboratories America
kzhang2@lbl.gov

CP2

An Lle Based Heterogeneous Metric Learning for Cross-Media Retrieval

With unstructured heterogeneous multimedia data such as texts, images being more and more widely used on the web, cross-media retrieval has become an increasingly important task. One of the key techniques in cross-media retrieval is how to compute distances or similarities among different types of media data. In this talk, we will introduce an LLE based heterogeneous metric learning method for cross-media retrieval.

Yi-Dong Shen, Peng Zhou, Liang Du
State Key Lab of Computer Science
Institute of Software, Chinese Academy of Sciences
ydshen@ios.ac.cn, zhoup@ios.ac.cn, duliang@ios.ac.cn

Mingyu Fan
Institute of Intelligent System and Decision,
Wenzhou University
fanmingyu@amss:ac:cn

CP3

Tensor Spectral Clustering for Partitioning Higher-Order Network Structures

Spectral methods represent an important class of tools for studying the structure of networks. However, they cannot explicitly encode higher-order network structures such as triangles, cycles, and feedforward loops. We propose a Tensor Spectral Clustering (TSC) framework that models higher-order network structures in a graph partitioning setting. TSC lets the user specify which higher-order network structures should be preserved by the network clustering. We demonstrate the efficacy of TSC on several synthetic and real-world examples.

Austin Benson
Stanford University
arbenson@stanford.edu

David F. Gleich
Purdue University
dgleich@purdue.edu

Jure Leskovec
Stanford University
jure@cs.stanford.edu

CP3

NetCodec: Community Detection from Individual Activities

The real social network and associated communities are often hidden under the declared friend or group lists in social networks. In this paper, we address the following question: *Could we simultaneously detect community structure and network infectivity among individuals from their time-stamped activities?* To this end, we parametrize the network infectivity in terms of individuals' participation in communities and the popularity of each individual. We show that this modeling approach has many benefits, both conceptually and experimentally. We utilize Bayesian variational inference to design NetCodec, an efficient inference algorithm which is verified with both synthetic and real world data sets. The experiments show that NetCodec can discover the underlying network infectivity and community structure more accurately than baseline method.

Long Q. Tran
University of Engineering and Technology
Vietnam National University at Hanoi
tqlong@vnu.edu.vn

Mehrdad Farajtabar, Le Song, Honguyan Zha
Georgia Tech
mehrdad.farajtabar@gmail.com, lsong@cc.gatech.edu, zha@cc.gatech.edu

CP3

Vertex Clustering of Augmented Graph Streams

We propose a graph stream clustering algorithm with a unified similarity measure on both structural and attribute properties of vertices, with each attribute being treated as a vertex. Unlike others, our approach does not require an input parameter for the number of clusters, instead, it dynamically creates new sketch-based clusters and period-

ically merges existing similar clusters. Experiments on two public datasets reveal the advantages of our approach in detecting vertex clusters in the graph stream.

Ryan Mcconville, Weiru Liu, Paul Miller
Queen's University Belfast
rmcconville07@qub.ac.uk, w.liu@qub.ac.uk, p.miller@qub.ac.uk

CP3

Efficient Algorithms for a Robust Modularity-Driven Clustering of Attributed Graphs

Today's applications store additional attribute information for each node in the graph. This attribute information may be contradicting with the structure, which raises a challenge for the simultaneous mining of both information sources. Thus, it is essential to be aware of irrelevant attributes and highly deviating attribute values of outlier nodes. In this work, we propose an efficient modularity-driven approach for parameter-free clustering of attributed graphs that is robust w.r.t. both irrelevant attributes and outliers.

Patricia Iglesias Sanchez, Emmanuel Müller, Uwe Leo Korn, Klemens Böhm, Andrea Kappes, Tanja Hartmann
Karlsruhe Institute of Technology
patricia.iglesias@kit.edu, emmanuel.mueller@kit.edu, uwe.korn@alumni.kit.edu, klemens.boehm@kit.edu, andrea.kappes@kit.edu, tanja.hartmann@kit.edu

Dorothea Wagner
Karlsruhe Institute of Technology
Institute of Theoretical Informatics
dorothea.wagner@kit.edu

CP3

Community Detection for Emerging Networks

In this paper, we want to detect communities for emerging networks. Community detection for emerging networks is very challenging as information in emerging networks is usually too sparse for traditional methods to calculate effective closeness scores among users and achieve good community detection results. Meanwhile, users nowadays usually join multiple social networks simultaneously, some of which are developed and can share common information with the emerging networks. Based on both link and attribution information across multiple networks, a new general closeness measure, intimacy, is introduced in this paper. With both micro and macro controls, an effective and efficient method, CAD (Cold stArt community Detector), is proposed to propagate information from developed network to calculate effective intimacy scores among users in emerging networks.

Jiawei Zhang
University of Illinois at Chicago
jzhan9@uic.edu

Philip S. Yu
University of Chicago
psyu@cs.uic.edu

CP4

Labeling Educational Content with Academic Learning Standards

Learning standards define the specific structure of an educational program. They contain a list of instructions specifying various skills that students should learn at different points during their learning progression. Manually identifying the appropriate learning standard instruction for learning content is time consuming and not scalable. In this paper, we address the problem of automatically labeling digital learning content with the learning standards. We demonstrate the usefulness of our approach on a collection of high school learning materials that were labeled by curriculum experts from a US school district according to a publicly available learning standard. The system developed has been deployed and is in use by the school district. To the best of our knowledge we are the first to attempt this novel task and develop such a system.

Danish Contractor
IBM Research
dcontrac@in.ibm.com

Kashyap Popat, Shajith Ikbal
IBM Research India
kaspopat@in.ibm.com, shajmoha@in.ibm.com

Sumit Negi, Bikram Sengupta, Mukesh Mohania
IBM Research
sumitneg@in.ibm.com, bsengupt@in.ibm.com, mkmukesh@in.ibm.com

CP4

Data Mining for Real Mining: A Robust Algorithm for Prospectivity Mapping with Uncertainties

Mineral prospectivity mapping is an application for machine learning which presents a series of practical difficulties. The goal is to learn the mapping function which can predict the existence of mineralization from a compilation of geoscience datasets. Challenges include sparse, imbalanced labels, varied label reliability, and a wide range in data uncertainty. To address this, an algorithm was developed based on TLS and SVM which incorporates both data and label uncertainty into the objective function.

Justin Granek
University of British Columbia
jgranek@eos.ubc.ca

Eldad Haber
Department of Mathematics
The University of British Columbia
haber@math.ubc.ca

CP4

Product Adoption Rate Prediction: A Multi-Factor View

Advances in commerce have enabled us to collect massive amounts of user consumption series, which can help predict the future product adoption for the merchants, thus enabling applications such as targeted marketing. However, previous works only aimed at predicting if one user will adopt this product or not; the problem of adoption rate (percentage of use) prediction for each user is still underexplored. This paper presents a comprehensive study for this problem. We first introduce a decision function to capture the change of users' product adoption rate. Then, we propose two solutions, the Generalized Adoption

Model (GAM) assuming all users are influenced equally by these factors and the Personalized Adoption Model (PAM) arguing each factor contribute differently among people. We further extend the PAM to a totally Bayesian model that can automatically learn all parameters. Finally, extensive experiments on two real-world datasets show the superiority of our proposed models.

Le Wu
University of Science and Technology of China
wule@mail.ustc.edu.cn

CP4

Combating Product Review Spam Campaigns via Multiple Heterogeneous Pairwise Features

Spam campaigns spotted in popular product review websites have attracted mounting attention from both industry and academia, where a group of online posters are hired to collaboratively craft deceptive reviews for some target products. The goal is to manipulate perceived reputations of the targets for their best interests. Compared to existing point-wise features extracted from individual reviewers or reviewer-groups, we find that pairwise features extracted from pairs of reviewers can be more robust to model the relationships among colluders since they, as the ingredients of spam campaigns, are correlated in nature. We explore multiple heterogeneous pairwise features in virtue of some collusion signals found in reviewers' rating behaviors and linguistic patterns. An unsupervised and intuitive colluder detecting framework has been proposed which can benefit from these pairwise features.

Chang Xu, Jie Zhang
School of Computer Engineering
Nanyang Technological University
xuch0007@e.ntu.edu.sg, zhangj@ntu.edu.sg

CP4

PatentCom: A Comparative View of Patent Document Retrieval

In this paper, we study the problem of comparing patent documents. We explore summarization strategies that generate comparative summaries to assist patent analysts in quickly reviewing any given patent document pairs. To this end, we present PatentCom, which first extracts discriminative terms from each document, and then connects the dots on a term co-occurrence graph. Extensive quantitative analysis and case studies on real world patent documents demonstrate the effectiveness of our proposed approach.

Longhui Zhang
florida international university
lzhan015@cs.fiu.edu

Lei Li, Chao Shen, Tao Li
Florida International University
lli003@cs.fiu.edu, cshen001@cs.fiu.edu, taoli@cs.fiu.edu

CP5

Binary Classifier Calibration Using a Bayesian Non-Parametric Approach

Learning probabilistic predictive models that are well calibrated is critical for many prediction and decision-making tasks in Data mining. This paper presents new Bayesian non-parametric methods for calibrating outputs of binary classification models. The advantage of these methods is that they are independent of the algorithm used to learn a predictive model, and they can be applied in a post-processing step. This makes them applicable to a wide variety of machine learning models.

Mahdi Pakdaman Naeini
Intelligent Systems Program
University of Pittsburgh
pakdaman@cs.pitt.edu

Gregory Cooper
Department of Biomedical Informatics
University of Pittsburgh
gfc@pitt.edu

Milos Hauskrecht
University of Pittsburgh
milos@cs.pitt.edu

CP5

A Bayesian Framework for Modeling Human Evaluations

Several situations that we come across in our daily lives involve some form of evaluation: a process where an evaluator judges a given item. Examples of such situations include a crowd-worker labeling an image or a student answering a multiple-choice question. Gaining insights into human evaluations is important for determining the quality of individual evaluators as well as identifying true labels of items. Here, we generalize the question of estimating the quality of individual evaluators, extending it to obtain diagnostic insights into how various evaluators label different kinds of items. We propose a series of increasingly powerful hierarchical Bayesian models with the goal of obtaining insights into the underlying evaluation process. We apply our framework to a wide range of real-world domains, and demonstrate that our approach can accurately predict evaluator decisions, diagnose types of mistakes evaluators tend to make, and infer true labels of items.

Himabindu Lakkaraju, Jure Leskovec
Stanford University
himalv@cs.stanford.edu, jure@cs.stanford.edu

Jon M. Kleinberg
Cornell University
Dept of Computer Science
kleinber@cs.cornell.edu

Sendhil Mullainathan
Harvard University
mullain@fas.harvard.edu

CP5

Cross-Modal Retrieval: A Pairwise Classification Approach

Content is increasingly available in multiple modalities (such as images, text, and video), each of which provides a different representation of some entity. The cross-modal retrieval problem is: given the representation of an entity in one modality, find its best representation in all other modalities. We propose a novel approach to this problem based on pairwise classification. The approach seamlessly applies to both the settings where ground-truth annotations for the entities are absent and present. In the latter

case, the approach considers both positive and *unlabelled* links that arise in standard cross-modal retrieval datasets. Empirical comparisons show improvements over state-of-the-art methods for cross-modal retrieval.

Aditya K. Menon
NICTA
aditya.menon@nicta.com.au

Didi Surian
The University of Sydney
NICTA
didi.surian@nicta.com.au

Sanjay Chawla
University of Sydney
sanjay.chawla@sydney.edu.au

CP5

Feature-Based Factorized Bilinear Similarity Model for Cold-Start Top-N Item Recommendation

The user personalized non-collaborative methods based on item features can be used to address this item cold-start problem. These methods rely on similari- ties between the target item and users previous preferred items. While computing similarities based on item features, these methods overlook the interactions among the features of the items and consider them independently. Modeling in- teractions among features can be helpful as some features, when considered together, provide a stronger signal on the relevance of an item when compared to case where features are considered independently. To address this important is- sue, in this work we introduce the Feature-based factorized Bilinear Similarity Model (FBSM), which learns factorized bilinear similarity model for Top-n recommendation of new item.

Mohit Sharma
University of Minnesota
sharm163@umn.edu

Jiayu Zhou, Junling Hu
Samsung
jiayu.zhou@samsung.com, junling.hu@samsung.com

George Karypis
University of Minnesota
karypis@umn.edu

CP5

Semi-Supervised Learning for Structured Regression on Partially Observed Attributed Graphs

In real–life applications a large fraction of observations is often missing which can severely limit the representational power of predictive models. In this paper we propose a Marginalized Gaussian CRF structured regression model for dealing with missing labels in partially observed temporal attributed graphs. The benefits of the new method are demonstrated on challenging application of predicting precipitation and on numerous synthetic graphs for various missingness mechanisms with up to 80% missing labels.

Jelena Z. Stojanovic
Temple University
jelena.stojanovic@temple.edu

Milos Jovanovic
University of Belgrade
milos.jovanovic@fon.bg.ac.rs

Djordje Gligorijevic, Zoran Obradovic
Temple University
gligorijevic@temple.edu, zoran.obradovic@temple.edu

CP6

Attacking Dbscan for Fun and Profit

Many security applications depend critically on clustering. However, we do not know of any clustering algorithms that were designed with an adversary in mind. An intelligent adversary might use this to subvert the security of the application. Already, adversaries use obfuscation and other techniques to alter the representation of their inputs to avoid detection. In this work, we investigate a more active attack, in which an adversary feeds carefully crafted inputs to the clustering analysis.

Jonathan Crussell, Philip Kegelmeyer
Sandia National Laboratories
jcrusse@sandia.gov, wpk@sandia.gov

CP6

Modeling Users' Adoption Behaviors with Social Selection and Influence

Two key factors that affect users' adoption behaviors are social selection and social influence. Understanding such factors underlying each behavior can potentially help web service providers gain much more insights into their users and improve predictive power. In this paper, we try to answer (1) How do the roles of selection and influence play in a user-level adoption? (2) Whether capturing those factors can benefit the modeling and prediction of users' adoption behaviors. We propose a probabilistic Latent Factors with Diffusion Model (LFDM) which explicitly considers both social selection and influence by projecting cascading processes into latent factor spaces, and develop an effective EM styled algorithm for estimating the proposed model. Finally we validate our methodology on three kinds of real world data sets.

Ziqi Liu
MOEKLINNS Lab, Department of Computer Science
Xi'an Jiaotong University, China
ziqilau@gmail.com

Fei Wang
Department of Computer Science and Engineering
University of Connecticut
fei_wang@uconn.edu

Qinghua Zheng
Department of Computer Science
Xi'an Jiaotong University, China
qhzheng@mail.xjtu.edu.cn

CP6

Health Insurance Market Risk Assessment: Covariate Shift and K-Anonymity

Health insurance companies prefer to enter new markets in which individuals likely to enroll in their plans have a low annual cost. When deciding which new markets to enter, health cost data for the new markets is unavailable to

them, but it is available for their own enrolled members. To address the problem of assessing risk in new markets, i.e., estimating the cost of likely enrollees, we pose a regression problem with demographic data as predictors combined with a novel three-population covariate shift. Since health data is protected by privacy laws, we cannot use the raw data of the insurance company's members directly for training the regression and covariate shift. Therefore, we also develop a novel method to achieve k-anonymity with the workload-driven quality of data distribution preservation achieved through dithered quantization and Rosenblatt's transformation. We illustrate the efficacy of the solution using real-world, publicly available data.

Dennis Wei, Karthikeyan Natesan Ramamurthy, Kush R. Varshney
Mathematical Sciences and Analytics Department
IBM Thomas J. Watson Research Center
dwei@us.ibm.com, knatesa@us.ibm.com, krvarshn@us.ibm.com

CP6

Result Integrity Verification of Outsourced Privacy-Preserving Frequent Itemset Mining

In the recently-emerged Data-Mining-as-a-Service ($DMaS$) paradigm, a client outsources her data and the data mining needs to a third party service provider. It raises a few security issues including privacy protection and result integrity verification. Most of the recent work studied these two issues separately. In this paper, we focus on the problem of result integrity verification of outsourced privacy-preserving frequent itemset mining. It is challenging to discover the incorrect results by the service provider's misbehaviors from the mining output that intends to be inaccurate due to privacy protection techniques. We design efficient approaches that can provide high probabilistic guarantee for both correctness and completeness of the frequent itemset mining results. Our experiment results show the efficiency and effectiveness of our approaches.

Wendy Hui Wang
Stevens Institute of Technology
hwang4@stevens.edu

Ruilin Liu
Department of Computer Science
Stevens Institute of Technology
rliu3@stevens.edu

CP6

Exploring the Impact of Dynamic Mutual Influence on Social Event Participation

Nowadays, it is commonly seen that an offline social event is organized through online social network services (SNS), in this way cyber strangers can be connected in physical world. While there are some preliminary studies on social event participation through SNS, they usually have more focus on the mining of event profiles and have less focus on the social relationships among target users. In particular, the importance of dynamic mutual influence among potential event participants has been largely ignored. In this paper, we develop a novel discriminant framework, which allows to integrate the dynamic mutual dependence of potential event participants into the discrimination process. Specifically, we formulate the group-oriented event participation problem as a variant two-stage discriminant framework to capture the users' preferences as well as their latent social connections. The experimental results on real-world data show that our method can effectively predict the event participation with a significant margin compared with several state-of-the-art baselines, which validates the hypothesis that dynamic mutual influence could play an important role in the decision-making process of social event participation.

Tong Xu
University of Science and Technology of China
tongxu@mail.ustc.edu.cn

Hao Zhong
Rutgers University
h.zhong31@rutgers.edu

Hengshu Zhu
Baidu Research - Big Data Lab
zhuhengshu@baidu.com

Hui Xiong
Rutgers, the State University of New Jersey
hxiong@rutgers.edu

Enhong Chen
University of Science and Technology of China
cheneh@ustc.edu.cn

Guannan Liu
Tsinghua University
guannliu@gmail.com

CP7

Fast Mining of a Network of Coevolving Time Series

Coevolving multiple time series are ubiquitous and naturally appear in a variety of high-impact applications, ranging from environmental monitoring, motion capture, to physiological signal in health care and many more. In many scenarios, the multiple time series data is often accompanied by some contextual information in the form of networks. In this paper, we refer to such multiple time series, together with its embedded network as *a network of coevolving time series*. In order to unveil the underlying patterns of a network of coevolving time series, we propose DCMF, a dynamic contextual matrix factorization algorithm. The key idea is to find the latent factor representation of the input time series and that of its embedded network simultaneously. Our experimental results on several real datasets demonstrate that our method (1) outperforms its competitors, especially when there are lots of missing values; and (2) enjoys a linear scalability w.r.t. the length of time series.

Yongjie Cai
The Graduate Center, City University of New York
yongjie207@gmail.com

Hanghang Tong
Arizona State University
hanghang.tong@asu.edu

Wei Fan
Big Data Labs - Baidu USA
fanwei03@baidu.com

Ping Ji
The Graduate Center, CUNY
pji@jjay.cuny.edu

CP7

Shapelet Ensemble for Multi-Dimensional Time Series

Time series shapelets are small subsequences that maximally differentiate classes of time series. Since the inception of shapelets, researchers have used shapelets for various data domains includ-ing anthropology and health care, and in the process suggested many efficient techniques for shapelet discovery. However, multi-dimensional time series data poses unique challenges to shapelet discovery that are yet to be solved. We show that an ensemble of shapelet-based decision trees on individual dimensions works better than shapelets defined over multiple dimensions. Generating a shapelet ensemble for multi-dimensional time series is computationally expensive. Most of the existing techniques prune shapelet candidates for speed. In this paper, we propose a novel technique for shapelet discovery that evaluates remaining candidates efficiently. Our algorithm uses a multi-length approximate index for time series data to efficiently find the nearest neighbors of the candidate shapelets. We employ a simple skipping technique for additional candidate pruning and a voting based technique to improve accuracy while retaining interpretability. Not only do we find a significant speed increase, our techniques enable us to efficiently discover shapelets on datasets with long time series such as hours of brain activity record-ings. We demonstrate our approach on a biomedical dataset and find significant differences between patients with schizophrenia and healthy controls.

Mustafa S. Cetin
University of New Mexico Department of Computer Science
musicet37@gmail.com

CP7

Efficient Online Relative Comparison Kernel Learning

Learning a kernel matrix from relative comparison feedback is an important problem with numerous applications. Existing methods that do so face significant scalability issues inhibiting their application to settings where a kernel is learned in an online and timely fashion. In this paper we propose a novel framework called **E**fficient online **R**elative comparison **K**ernel **LE**arning (ERKLE) for efficiently learning the similarity of a large set of objects in an online manner. We take advantage of the sparse and low-rank properties of the stochastic gradient with respect to a single comparison to efficiently restrict the kernel to lie in the space of positive semidefinite matrices. In addition, we derive a passive-aggressive update for minimally satisfying new relative comparisons as to not disrupt the influence of previously obtained comparisons. Experimentally, we demonstrate a considerable improvement in speed while obtaining improved or comparable accuracy compared to current methods.

Eric Heim
University of Pittsburgh
eric@cs.pitt.edu

Matthew Berger, Lee Seversky
Air Force Research Laboratory, Information Directorate
matthew.berger.1@us.af.mil, lee.seversky@us.af.mil

Milos Hauskrecht
University of Pittsburgh
milos@cs.pitt.edu

CP7

***Cheetah:* Fast Graph Kernel Tracking on Dynamic Graphs**

Graph kernels provide an expressive approach to measuring the similarity of two graphs, and are key building blocks behind many real-world applications, such as bioinformatics, brain science and social networks. However, current methods for computing graph kernels assume the input graphs are static, which is often not the case in reality. It is highly desirable to track the graph kernels on dynamic graphs evolving over time in a timely manner. In this paper, we propose a family of *Cheetah* algorithms to deal with the challenge. *Cheetah* leverages the low rank structure of graph updates and incrementally updates the eigen-decomposition or SVD of the adjacency matrices of graphs. Experimental evaluations on real world graphs validate our algorithms (1) are significantly faster than alternatives with high accuracy and (b) scale sub-linearly.

Liangyue Li, Hanghang Tong
Arizona State University
liangyue@asu.edu, hanghang.tong@asu.edu

Yanghua Xiao
Fudan University
shawyh@fudan.edu.cn

Wei Fan
Big Data Labs - Baidu USA
fanwei03@baidu.com

CP7

On the Non-Trivial Generalization of Dynamic Time Warping to the Multi-Dimensional Case

We show the two most commonly used multi-dimensional DTW methods, dependent or independent warping, produce different classification results. We present a simple, principled rule which can be used on a case-by-case basis to predict which of the two methods we should pick. Our method allows us to ensure that classification results are at least as accurate as the better of the two rival methods, and in many cases, our method is strictly more accurate.

Mohammad Shokoohi-Yekta
University of California, Riverside
mshok002@cs.ucr.edu

Jun Wang
University of Texas, Dallas
wangjun@utdallas.edu

Eamonn Keogh
University of California, Riverside
eamonn@cs.ucr.edu

CP8

Low Rank Representation on Riemannian Manifold

of Symmetric Positive Definite Matrices

In many computer vision applications, data often originate from a manifold, which is equipped with some Riemannian geometry. In this case, the existing Low Rank Representation (LRR) becomes inappropriate for modeling and incorporating the intrinsic geometry of the manifold that is potentially important and critical to applications. In this paper, we generalize the LRR over the Euclidean space to the LRR model over a specific Rimannian manifold—the manifold of symmetric positive matrices (SPD).

Yifan Fu
Charles Sturt University
yfu@csu.edu.au

Junbin Gao
Charles Sturt University, Australia
jbgao@csu.edu.au

Xia Hong
University of Reading
x.hong@reading.ac.uk

David Tien
Charles Sturt University
dtien@csu.edu.au

CP8

Getting to Know the Unknown Unknowns: Destructive-Noise Resistant Boolean Matrix Factorization

Many real-world binary datasets exhibit mostly destructive noise. This is because failure to observe something that exists is more likely than reporting something that does not. Existing methods are not robust against destructive noise, however. To handle this phenomenon we use the MDL principle with Boolean matrix factorization. Our algorithm, Nassau, finds Boolean factorization with short description length. This allows it to find good and intuitive factorizations from both synthetic and real-world datasets.

Sanjar Karaev
Max Planck Institute for Informatics
skaraev@mpi-inf.mpg.de

Pauli Miettinen
Max-Plank Institute for Informatics
Saarbruecken, Germany
pmiettin@mpi-inf.mpg.de

Jilles Vreeken
Max Planck Institute for Informatics
Saarland University
jilles@mpi-inf.mpg.de

CP8

Near-Separable Non-Negative Matrix Factorization with ℓ_1 and Bregman Loss Functions

Recently, a family of tractable NMF algorithms have been proposed under the assumption that the data matrix satisfies a separability condition Donoho & Stodden (2003); Arora et al. (2012). Geometrically, this condition reformulates the NMF problem as that of finding the extreme rays of the conical hull of a finite set of vectors. In this paper, we develop several extensions of the conical hull procedures of Kumar et al. (2013) for robust (ℓ_1) approximations and Bregman divergences. Our methods inherit all the advantages of Kumar et al. (2013) including scalability and noise-tolerance. We show that on foreground-background separation problems in computer vision, robust near-separable NMFs match the performance of Robust PCA, considered state of the art on these problems, with an order of magnitude faster training time. We also demonstrate applications in exemplar selection settings.

Abhishek Kumar
IBM Research
IBM Research
abhishek@umiacs.umd.edu

Vikas Sindhwani
Google Research
sindhwani@google.com

CP8

Personalized TV Recommendation with Mixture Probabilistic Matrix Factorization

The rapid development of smart TV leads to a great demand of building personalized TV recommender system. While different methods have been proposed, most of them neglect the mixture of watching groups behind an individual TV. To this end, we propose a Mixture Probabilistic Matrix Factorization model to learn the program preferences of televisions, which assumes that the preference of a given television can be regarded as the mixed preference of different watching groups.

Huayu Li
University of North Carolina at Charlotte
hli38@uncc.edu

Hengshu Zhu
Baidu Research - Big Data Lab
zhuhengshu@baidu.com

Yong Ge
UNC Charlotte
yong.ge@uncc.edu

Yanjie Fu
Rutgers University
yanjie.fu@rutgers.edu

Yuan Ge
Anhui Polytechnic University
ygetoby@mail.ustc.edu.cn

CP8

Convex Matrix Completion: A Trace-Ball Optimization Perspective

In this presentation, we will first introduce our *trace-ball optimization* method for convex trace bounding matrix completion problem, which can creatively change the original trace norm constraint into the problem of low-rank matrix factorization. Then we study on the properties about the free parameter of the convex trace bounding matrix completion problem.

Guangxiang Zeng
University of Science and Technology of China

zgx@mail.ustc.edu.cn

Ping Luo
ICT, CAS
luop@ict.ac.cn

Enhong Chen
University of Science and Technology of China
cheneh@ustc.edu.cn

Hui Xiong
Rutgers, the State University of New Jersey
hxiong@rutgers.edu

Hengshu Zhu
Baidu Research - Big Data Lab
zhuhengshu@baidu.com

Qi Liu
University of Science and Technology of China
qiliuql@ustc.edu.cn

CP9

Legislative Prediction with Dual Uncertainty Minimization from Heterogeneous Information

In this paper, we present a novel prediction model that maximizes the usage of publicly accessible heterogeneous data, i.e., bill text and lawmakers' profile data, to carry out effective legislative prediction. In particular, we propose to design a probabilistic prediction model which achieves high consistency with past vote records while ensuring the minimum uncertainty of the vote prediction reflecting the firm legal ground often held by the lawmakers. In addition, the proposed legislative prediction model enjoys the following properties: inductive and analytical solution, abilities to deal with the prediction on new bills and new legislators, and robustness to the missing vote issue. We conduct extensive empirical study using the real legislative data. The experimental results clearly corroborate that the proposed method provides superior prediction accuracy with visible performance gain.

Yu Cheng
Northwestern University
IBM T.J. Watson Research Center
chengyu05@gmail.com

Ankit Agrawal
Northwestern University
ankitag@eecs.northwestern.edu

Huan Liu
Arizona State University
ankitag@eecs.northwestern.edu

Alok Choudhary
Dept. of Electrical Engineering and Computer Science
Northwestern University, Evanston, USA
choudhar@eecs.northwestern.edu

CP9

Gin: A Clustering Model for Capturing Dual Heterogeneity in Networked Data

Networked data often consists of interconnected multi-typed nodes and links. A common assumption behind such heterogeneity is the shared clustering structure. However, existing network clustering approaches oversimplify the heterogeneity by either treating nodes or links in a homogeneous fashion, resulting in massive loss of information. In addition, these studies are more or less restricted to specific network schemas or applications, losing generality. In this paper, we introduce a flexible model to explain the process of forming heterogeneous links based on shared clustering information of heterogeneous nodes. Specifically, we categorize the link generation process into binary and weighted cases and model them respectively. We show these two cases can be seamlessly integrated into a unified model.

Jialu Liu
University of Illinois at Urbana-Champaign
jliu64@illinois.edu

Chi Wang
Microsoft Research Redmond
chiw@microsoft.com

Jing Gao
University at Buffalo
jing@buffalo.edu

Quanquan Gu
University of Virginia
qg5w@virginia.edu

Charu C. Aggarwal
IBM T. J. Watson Research Center
charu@us.ibm.com

Lance Kaplan
U.S. Army Research Laboratory
lance.m.kaplan@us.army.mil

Jiawei Han
UIUC
hanj@illinois.edu

CP9

SourceSeer: Forecasting Rare Disease Outbreaks Using Multiple Data Sources

Rapidly increasing volumes of news feeds from diverse data sources are extremely valuable in forecasting rare disease outbreaks. We present SourceSeer, a framework that combines spatio-temporal topic models with source-based anomaly detection techniques to effectively forecast rare disease outbreaks. SourceSeer discovers the location focus sources treating them as experts with varying degrees of authoritativeness. We evaluate the performance of SourceSeer on forecasting hantavirus outbreaks in Latin America and demonstrate increased accuracy compared to several baselines.

Theodoros Rekatsinas
University of Maryland
thodrek@cs.umd.edu

Saurav Ghosh
Virginia Tech
sauravcsvt@cs.vt.edu

Sumiko Mekaru
Children's Hopsital Boston, MA, USA

sumiko.mekaru@childrens.harvard.edu

Elaine Nsoesie
Boston Children's Hospital
elaine.nsoesie@childrens.harvard.edu

John Brownstein
Children's Hospital Boston, MA, USA
john.brownstein@childrens.harvard.edu

Lise Getoor
University of California, Santa Cruz
getoor@soe.ucsc.edu

Naren Ramakrishnan
Computer Science
Virginia Tech
naren@cs.vt.edu

CP9

Plums: Predicting Links Using Multiple Sources

Link prediction is an important problem in online networks, for recommending friends and collaborators. We build a robust and effective classifier for link prediction using multiple auxiliary networks. Competing methods are mostly feature-based and construct a large number of network-level features to make the prediction more effective. We develop a supervised random walk model, without explicit feature construction, personalized to each user based on past accept and reject behavior. Our approach consistently outperforms popular baselines.

Karthik Subbian, Arindam Banerjee
University of Minnesota
mailtosuka@gmail.com, banerjee@cs.umn.edu

Sugato Basu
Google Research
sugato@google.com

CP9

Believe It Today Or Tomorrow? Detecting Untrustworthy Information from Dynamic Multi-Source Data

In this paper, we conduct trustworthiness analysis from a novel perspective of correlating and comparing multiple sources that describe the same set of items. Different from existing work, we recognize the importance of time dimension in modeling the commonalities. We represent dynamic sparse multi-source data as tensors and develop both offline and incremental tensor factorization approaches to capture the common patterns across sources. Results demonstrate the advantages of the proposed approach in detecting untrustworthy information.

Houping Xiao, Yaliang Li
SUNY Buffalo
houpingx@buffalo.edu, yaliangl@buffalo.edu

Jing Gao
University at Buffalo
jing@buffalo.edu

Fei Wang
Department of Computer Science and Engineering
University of Connecticut
fei_wang@uconn.edu

Liang Ge
Google
lge@google.com

Wei Fan
Big Data Labs - Baidu USA
fanwei03@baidu.com

Long Vu
IBM T.J. Watson Research Center
lhvu@us.ibm.com

Deepak Turaga
IBM Research
turaga@us.ibm.com

CP10

Clustering and Ranking in Heterogeneous Information Networks via Gamma-Poisson Model

We propose a probabilistic generative model that simultaneously achieves clustering and ranking on a heterogeneous network that can follow arbitrary schema, where the edges from different types are sampled from a Poisson distribution with the parameters determined by the ranking scores of the nodes in each cluster. A variational Bayesian inference method is proposed to learn these parameters, which can be used to output ranking and clusters simultaneously.

Junxiang Chen, Wei Dai, Yizhou Sun, Jennifer Dy
Northeastern University
jchen@ece.neu.edu, wei@ece.neu.edu, yzsun@ccs.neu.edu, jdy@ece.neu.edu

CP10

A Devide-and-Conquer Algorithm for Betweenness Centrality

The problem of efficiently computing the betweenness centrality of nodes has been researched extensively. To date, the best known exact and centralized algorithm for this task is an algorithm proposed in 2001 by Brandes. The contribution of our paper is `Brandes++`, an algorithm for exact efficient computation of betweenness centrality. The crux of our algorithm is that we create a sketch of the graph, that we call the skeleton, by replacing subgraphs with simpler graph structures. Depending on the underlying graph structure, using this skeleton and by keeping appropriate summaries `Brandes++` we can achieve significantly low running times in our computations. Extensive experimental evaluation on real life datasets demonstrate the efficacy of our algorithm for different types of graphs. We release our code for benefit of the research community.

Dora Erdos
Boston University
edori@bu.edu

Vatche Ishakian
IBM T J Watson Research Center
vishaki@us.ibm.com

Azer Bestavros, Evimaria Terzi
Boston University

best@cs.bu.edu, evimaria@cs.bu.edu

CP10

Frameworks to Encode User Preferences for Inferring Topic-Sensitive Information Networks

In online social networks, we often easily observe the time when each user receives a message, yet the users connections empowering the message diffusion remain hidden. This talk addresses the problem of uncovering the hidden diffusion network from the traces of disseminated messages. We introduce two principled methods: ***Weighted Topic Cascade*** (WTC) and ***Preference-enhanced Topic Cascade*** (PTC), which incorporate both user preferences and the topic distribution of messages to assist the network inference process.

Qingbo Hu, Sihong Xie, Shuyang Lin
University of Illinois at Chicago
qhu5@uic.edu, sxie6@uic.edu, slin38@uic.edu

Wei Fan
Big Data Labs - Baidu USA
fanwei03@baidu.com

Philip Yu
University of Illinois at Chicago
psyu@uic.edu

CP10

Hidden Hazards: Finding Missing Nodes in Large Graph Epidemics

Given a noisy or sampled snapshot of an infection in a large graph, can we automatically and reliably recover the truly infected yet somehow missed nodes? And, what about the seeds, the nodes from which the infection started to spread? These are important questions in diverse contexts, ranging from epidemiology to social media. In this paper, we address the problem of simultaneously recovering the missing infections and the source nodes of the epidemic given noisy data. We formulate the problem by the Minimum Description Length principle, and propose NetFill, an efficient algorithm that automatically and highly accurately identifies the number and identities of both missing nodes and the infection seed nodes. Experimental evaluation on synthetic and real datasets, including using data from information cascades over 96 million blog posts and news articles, shows that our method outperforms other baselines, scales near-linearly, and is highly effective in recovering missing nodes and sources.

Aditya Prakash
Virginia Tech
badityap@cs.vt.edu

CP10

Rare Class Detection in Networks

Many networks are content-rich, and the content-rich nature of such networks can be leveraged to compensate for the lack of structural connectivity among rare class nodes. While content-centric and semi-supervised methods have been used earlier in the context of paucity of labeled data, the rare class scenario has not been investigated in this context. This paper will present a spectral approach for rare-class detection, which uses a distance-preserving transform, in order to combine the structure and content in the network.

Karthik Subbian
University of Minnesota
mailtosuka@gmail.com

Charu C. Aggarwal
IBM T. J. Watson Research Center
charu@us.ibm.com

Jaideep Srivastava
Qatar Computing Research Institute
jsrivastava@qf.org.qa

Vipin Kumar
University of Minnesota
kumar@cs.umn.edu

CP11

An ADMM Algorithm for Clustering Partially Observed Networks

To detect underlying cluster structure in a network, we propose a convex model, tighter than the robust PCA formulation, to decompose a partially observed adjacency matrix into low-rank and sparse components such that the low-rank component encodes the cluster structure. We devise an ADMM algorithm to solve this problem; and compare it with Louvain method, which maximizes modularity. Numerical results show that our method outperforms Louvain method when variance among cluster sizes is high.

Necdet S. Aybat
Columbia University
nsa10@psu.edu

Sahar Zarmehri, Soundar Kumara
Penn State University
sxz155@psu.edu, skumara@psu.edu

CP11

A Distributed Frank-Wolfe Algorithm for Communication-Efficient Sparse Learning

Learning sparse combinations is a frequent theme in machine learning. In this paper, we study its associated optimization problem in the distributed setting where the elements to be combined are not centrally located but spread over a network. We address the key challenges of balancing communication costs and optimization errors. To this end, we propose a distributed Frank-Wolfe (dFW) algorithm. We obtain theoretical guarantees on the optimization error e and communication cost that do not depend on the total number of combining elements. We further show that the communication cost of dFW is optimal by deriving a lower-bound on the communication cost required to construct an e-approximate solution. We validate our theoretical analysis with empirical studies on synthetic and real-world data, which demonstrate that dFW outperforms both baselines and competing methods. We also study the performance of dFW when the conditions of our analysis are relaxed, and show that dFW is fairly robust.

Aurélien Bellet
Télécom ParisTech
bellet@usc.edu

Yingyu Liang
Princeton University
yingyul@cs.princeton.edu

Alireza Bagheri Garakani
University of Southern California
bagherig@usc.edu

Maria-Florina Balcan
Carnegie Mellon University
ninamf@cs.cmu.edu

Fei Sha
University of Southern California
feisha@usc.edu

CP11

Exceptional Model Mining with Tree-Constrained Gradient Ascent

We propose a new algorithm called tree-constrained gradient ascent (TCGA) that exploits information on the contribution of individual records to subgroup quality and at the same time guarantees that the subgroup extension can be described concisely in the pattern language. To this end we generalize the notion of a subgroup to that of a *soft* subgroup. Hereby the quality measure is made differentiable, and the quality of a subgroup extension is then optimized numerically using a form of constrained gradient ascent. The constraint is constructed simultaneously with the optimization in order to ensure that subgroups can be described in the pattern language, and at the same time hinder the numerical optimization as little as possible. Subsequently we apply a post-processing step to further simplify and generalize the patterns found by the algorithm.

Ad Feelders, Thomas Krak
Universiteit Utrecht
A.J.Feelders@uu.nl, t.e.krak@uu.nl

CP11

Scaling Log-Linear Analysis to Datasets with Thousands of Variables

The primary statistical approach to association discovery between variables is log-linear analysis. Tackling high-dimensional datasets remained impossible, because every step of the search had a quadratic complexity with the number of variables. We show that only a very small subset of elements has to be considered at each step of the search, which makes it possible to perform log-linear analysis 4 orders of magnitude faster, and thus to target datasets with thousands of variables.

Francois Petitjean, Geoffrey Webb
Monash University
francois.petitjean@monash.edu, geoff.webb@monash.edu

CP11

Dropout Training of Matrix Factorization and Autoencoder for Link Prediction in Sparse Graphs

Matrix factorization (MF) and Autoencoder (AE) are among the most successful approaches of unsupervised learning. While MF based models have been extensively exploited in the graph modeling and link prediction literature, the AE family has not gained much attention. In this paper we investigate both MF and AE's application to the link prediction problem in sparse graphs. We show the connection between AE and MF from the perspective of multiview learning, and further propose MF+AE: a model training MF and AE jointly with shared parameters. We apply dropout to training both the MF and AE parts, and show that it can significantly prevent overfitting by acting as an adaptive regularization. We conduct experiments on six real world sparse graph datasets, and show that MF+AE consistently outperforms the competing methods, especially on datasets that demonstrate strong non-cohesive structures.

Shuangfei Zhai
State University of New York at Binghamton
szhai2@binghamton.edu

CP12

Active Multi-Task Learning Via Bandits

In this paper, we propose a new active multi-task learning paradigm, which selectively samples effective instances for multi-task learning. Inspired by the multi-armed bandits, which can balance the trade-off between the exploitation and exploration, we introduce a new active learning strategy and cast the selection procedure as a bandit framework. We consider both the risk of multi-task learner and the corresponding confidence bounds and our selection tries to balance this trade-off. Our proposed method is a sequential algorithm, which at each round maintains a sampling distribution on the pool of data, queries the label for an instance according to this distribution and updates the distribution based on the newly trained multi-task learner.

Meng Fang
University of Technology Sydney
Meng.Fang@student.uts.edu.au

Dacheng Tao
University of Technology, Sydney
dacheng.tao@uts.edu.au

CP12

Hierarchical Active Transfer Learning

We describe a unified active transfer learning framework called Hierarchical Active Transfer Learning (HATL). HATL exploits cluster structure shared between domains to perform transfer learning by imputing labels for unlabeled points and to generate label queries during active learning. We derive an upper bound on HATL's error when inferring labels for unlabeled data and present synthetic data results that confirm our analysis. Finally, we demonstrate HATL's empirical effectiveness on a benchmark sentiment classification data set.

David Kale
University of Southern California
1981
dkale@usc.edu

Marjan Ghazvininejad, Anil Ramakrishna
University of Southern California
ghazvini@isi.edu, akramakr@usc.edu

Jingrui He
Arizona State University
jingrui.he@asu.edu

Yan Liu
University of Southern California
yanliu.cs@usc.edu

CP12

FORMULA: FactORized MUlti-task LeArning for Task Discovery in Personalized Medical Models

Medical predictive modeling is challenging due to the heterogeneity of patients. In order to build effective predictive models we need to address such heterogeneous nature and allow patients to have their own personalized models. However, building a personalized model for each patient is computationally expensive and susceptible to overfitting. To address these challenges, we propose a novel approach called FactORized MUlti-task LeArning model (FORMULA). The personalized models are assumed to share a set of low-rank base models and learned by a sparse multi-task learning method. FORMULA is designed to simultaneously learn the base models as well as the personalized model of each patient, where the latter is a linear combination of the base models. We have performed extensive experiments to evaluate the proposed approach on a real medical data set. The proposed approach delivered superior predictive performance with useful medical insights.

Jianpeng Xu
Computer Science and Engineering Department
Michigan State University
xujianpe@msu.edu

Jiayu Zhou
Arizona State University
Jiayu.Zhou@asu.edu

Pang-Ning Tan
Michigan State University
ptan@cse.msu.edu

CP12

Faster Jobs in Distributed Data Processing Using Multi-Task Learning

Despite existing mitigation techniques, stragglers in distributed processing frameworks can significantly extend job completion times on production clusters, leading to increased costs. Proactive techniques (Wrangler) improve task scheduling by using predictive models. To capture inherent variability, separate models are built for every node and workload, requiring the time consuming collection of training data and limiting generalization to new nodes and workloads. Since predictors for similar nodes or workloads are likely to be similar and can share information, we extend the multi-task learning formulation of [**?**] to capture this group structure. Compared to Wrangler, our formulation improves straggler prediction accuracy by 7%, reduces job completion times by up to 59%, and for comparable accuracy requires only a 6th of the training data and thus much less training time. Our formulation also generalizes much better to tasks with insufficient data.

Neeraja J. Yadwadkar, Bharath Hariharan, Joseph Gonzalez, Randy Katz
University of California, Berkeley
neerajay@eecs.berkeley.edu, bharath2@cs.berkeley.edu, jegonzal@cs.berkeley.edu, randy@cs.berkeley.edu

CP12

Learning Complex Rare Categories with Dual Heterogeneity

We study complex rare categories with both task and view heterogeneity, and propose a novel optimization framework. It introduces a boundary characterization metric to capture the sharp changes in density near the boundary of the rare categories in the feature space, and constructs a graph-based model to leverage both task and view heterogeneity. We present an effective algorithm to solve this framework, analyze its performance from various aspects, and demonstrate its effectiveness on various datasets.

Pei Yang
Arizona State University
cs.pyang@gmail.com

Jingrui He
Stevens Institute of Technology
Computer Science Department
jingrui.he@gmail.com

Jia-Yu Pan
Google Inc.
jiayu.pan@gmail.com

CP13

Tracking Events Using Time-Dependent Hierarchical Dirichlet Tree Model

Timeline Generation, through generating news timelines from the massive data of news corpus, aims at providing readers with summaries about the evolvement of an event. In this paper, we develop a novel time-dependent Hierarchical Dirichlet Tree Model (tHDT) for timeline generation. Our model can aptly detect different levels of topic information in corpus and the structure is further used for sentence selection. Based on the topic distribution mined from tHDT, sentences are selected through an overall consideration of relevance, coherence and coverage. We develop experimental systems to compare different rival algorithms on 8 long-term events of public concern. The performance comparison demonstrates the effectiveness of our proposed model in terms of ROUGE metrics.

Rumeng Li
Peking University
alicerumeng@foxmail.com

Tao Wang
Wuhan University
whwtao@gmail.com

Xun Wang
Peking University
wangxun.pku@gmail.com

CP13

Selecting Social Media Responses to News: A Convex Framework Based On Data Reconstruction

With the explosive growth of social media, it has gained significantly increasing attention from both journalists and their readership in recent years by enhancing the reading

experience with its timeliness, high participation, interactivity, etc. On the other hand, the popularity of social media services such as twitter also leads to the challenge of information overload by generating thousands of messages (tweets) for each article of hot news, which will be overwhelming for readers. In this paper, we address the problem of selecting a representative subset of responses to news in order to deliver the most important information. We consider different criteria regarding the importance of the selected subset, and treat the problem from the data reconstruction perspective with concerns for both quality and generalizability of the selection. The intuition behind our work is that a good selection should be relevant from two levels: i) at the message level, it brings readers new information as much as possible or generalizes other people's opinions comprehensively; ii) at the text level, it is able to reconstruct the corpus. Specifically, the task of selecting responses to news can be formulated as a convex optimization problem where sparse non-negative weights are introduced for all the responses indicating whether they are selected or not. Several gradient based optimization and step size selection methods are also investigated in this paper to achieve a faster rate of convergence. More importantly, the proposed framework evaluates the utility of a set of responses jointly and therefore is able to reduce redundancy of the selected responses. We evaluate our approach on real-world data obtained from Twitter, and the results demonstrate superior performance over the state of the art in both accuracy and generalizability.

Linli Xu
University of Science and Technology of China
linlixu@ustc.edu.cn

CP14

Fast Eigen-Functions Tracking on Dynamic Graphs

Many important graph parameters can be expressed as eigen-functions of its adjacency matrix. Examples include epidemic threshold, graph robustness, etc. It is often of key importance to accurately monitor these parameters. However, most, if not all, of the existing algorithms computing these measures assume that the input graph is static, despite the fact that almost all real graphs are evolving over time. In this paper, we propose two online algorithms to track the eigen-functions of a dynamic graph with linear complexity wrt the number of nodes and number of changed edges in the graph. The key idea is to leverage matrix perturbation theory to efficiently update the top eigen-pairs of the underlying graph without re-computing them from scratch at each time stamp. Experiment results demonstrate that our methods can reach up to 20× speedup with precision more than 80% for fairly long period of time.

Chen Chen
Arizona State Univerisity
chenannie45@gmail.com

Hanghang Tong
Arizona State University
hanghang.tong@asu.edu

CP14

Approximation Algorithms for Reducing the Spectral Radius to Control Epidemic Spread

For several models of epidemic spread on networks (e.g., the 'flu-like' SIS model), it has been shown that an epidemic dies out quickly if the spectral radius of the graph is below a certain threshold that depends on the model parameters. This motivates a epidemic containment strategy to control epidemic spread by reducing the spectral radius of the underlying network. We develop a suite of provable approximation algorithms for reducing the spectral radius by removing the minimum cost set of edges or nodes, with different time and quality tradeoffs.

Sudip Saha
Virginia Polytechnic Institute and State University
ssaha@vbi.vt.edu

Abhijin Adiga
Virginia Tech
abhijin@vbi.vt.edu

B. Aditya Prakash
CS, VT
badityap@cs.vt.edu

Anil Vullikanti
Dept. of Computer Science, and Virginia Bioinformatics Inst.
Virginia Tech
akumar@vbi.vt.edu

CP15

Polyglot-Ner: Massive Multilingual Named Entity Recognition

The increasing diversity of languages used on the web introduces a new level of complexity to Information Retrieval (IR) systems. We can no longer assume that textual content is written in one language or even the same language family. In this paper, we demonstrate how to build massive multilingual annotators with minimal human expertise and intervention. We describe a system that builds Named Entity Recognition (NER) annotators for **40 major languages** using WIKIPEDIA and FREEBASE. The novelty of approach lies therein - using only language agnostic techniques, while achieving competitive performance. Our evaluation is two fold: First, we demonstrate the system performance on human annotated datasets. Second, for languages where no gold-standard benchmarks are available, we propose a new method, *distant evaluation*, based on statistical machine translation.

Rami Al-Rfou
Stony Brook University
Professor
ralrfou@cs.stonybrook.edu

CP15

Online Resource Allocation with Structured Diversification

A key consideration in online resource allocation problems is some notion of risk, and suitable ways to alleviate risk. Often, the risk is structured so that groups of assets are exposed to similar risks. We present a formulation for online resource allocation with structured diversification as a constrained online convex optimization problem. The key novel component of our formulation is a constraint on the $L_{(\infty,1)}$ group norm of the resource allocation vector.

Nicholas A. Johnson, Arindam Banerjee
University of Minnesota

njohnson@cs.umn.edu, banerjee@cs.umn.edu

CP15

Towards Permission Request Prediction on Mobile Apps Via Structure Feature Learning

The popularity of mobile apps has posed severe privacy risks to users because many permissions are over-claimed. In this work, we explore the techniques that can automatically predict the permission requests of a new mobile app based on its functionality and textual description information, which can help users to be aware of the privacy risks of mobile apps. Our framework formalizes the permission prediction problem as a multi-label learning problem, where a regularized structure feature learning framework is utilized to automatically capture the relations among textual descriptions, permissions, and app category. We evaluate our approach on 173 permission requests from 11,067 mobile apps across 30 categories. Extensive experiment results indicate that our method consistently provides better performance (3%-5% performance improvement in terms of F1 score), when compared to the other state-of-the-art methods.

Deguang Kong
University of Texas, Arlington
doogkong@gmail.com

Hongxia Jin
Samsung Research America
hongxia.jin@samsung.com

CP15

Propagation-Based Sentiment Analysis for Microblogging Data

The explosive popularity of microblogging services encourages more and more online users to share their opinions, and sentiment analysis on such opinion-rich resources has been proven to be an effective way to understand public opinions. On the one hand, the brevity and informality of microblogging data plus its wide variety and rapid evolution of language in microblogging pose new challenges to the vast majority of existing methods. On the other hand, microblogging texts contain various types of emotional signals strongly associated with their sentiment polarity, which brings about new opportunities for sentiment analysis. In this paper, we investigate propagation-based sentiment analysis for microblogging data. In particular, we provide a propagating process to incorporate various types of emotional signals in microblogging data into a coherent model, and propose a novel sentiment analysis framework PSA.

Jiliang Tang
Arizona State University
ARIZONA STATE UNIVERISTY
Jiliang.Tang@asu.edu

Chikashi Nobata, Anlei Dong, Yi Chang
Yahoo Labs
chikashi@yahoo-inc.com, anlei@yahoo-inc.com,
yichang@yahoo-inc.com

Huan Liu
Arizona State University
huan.liu@asu.edu

CP16

Estimating Ad Impact on Clicker Conversions for Causal Attribution: A Potential Outcomes Approach

We analyze the causal effect of online ads on the conversion probability of the users who click on the ad (clickers). We show that designing a randomized experiment to find this effect is infeasible, and propose a method to find the local effect on the clicker conversions. This method is developed in the Potential Outcomes causal model, via Principal Stratification to model non-ignorable post-treatment variables. Based on two large-scale randomized experiments, performed for 7.16 million and 22.7 million users, a pessimistic analysis shows a minimum increase of the effect on the clicker conversion probability of 75% with respect to the non-clickers. This finding contradicts the belief that clicks are not indicative of ad success. We find that a larger number of converting users is attributed to the overall campaign than those based on the click-to-conversion (C2C) business model. This evidence challenges the well-accepted belief that C2C model over-estimates the campaign value.

Joel Barajas
University of California, Santa Cruz
jbarajas@soe.ucsc.edu

Ram Akella
University of California, Berkeley
akella@ischool.berkeley.edu

Aaron Flores, Marius Holtan
AOL Research, Palo Alto, CA, USA
aaron.flores@teamaol.com, marius.holtan@teamaol.com

CP16

On Influential Nodes Tracking in Dynamic Social Networks

We explore the *Influential Node Tracking* problem which focuses on tracking a set of influential nodes that keeps maximizing the influence as the network structure evolves. Utilizing the smoothness of the evolution of the network, we implement node replacement to improve the influence coverage from the previous seed set instead of constructing the seed set from the ground. Experiments show that our method achieves better performance in terms of both influence coverage and running time.

Xiaodong Chen, Guojie Song
Peking University
chenxd@pku.edu.cn, gjsong@pku.edu.cn

Xinran He
University of Southern California
xinranhe@usc.edu

Kunqing Xie
Peking Universiry
kunqing@cis.pku.edu.cn

CP16

Principled Neuro-Functional Connectivity Discov-

ery

How can we reverse-engineer the brain connectivity, given the input stimulus, and the corresponding brain-activity measurements, for several experiments? We show how to solve the problem in a principled way, modeling the brain as a linear dynamical system (LDS), and solving the resulting "system identification' problem after imposing sparsity and non-negativity constraints on the appropriate matrices. These are reasonable assumptions in some applications, including magnetoencephalography (MEG). There are three contributions: (a) *Proof*: We prove that this simple condition resolves the ambiguity of similarity transformation in the LDS identification problem; (b) *Algorithm*: we propose an effective algorithm which further induces sparse connectivity in a principled way; and (c) *Validation*: our experiments on semi-synthetic (C. elegans), as well as real MEG data, show that our method recovers the neural connectivity, and it leads to interpretable results.

Kejun Huang, Nicholas Sidiropoulos
University of Minnesota
huang663@umn.edu, nikos@ece.umn.edu

Evangelos Papalexakis
CMU
epapalex@cs.cmu.edu

Christos Faloutsos
Carnegie Mellon University
christos@cs.cmu.edu

Partha Talukdar
Indian Institute of Science
ppt@serc.iisc.in

Tom Mitchell
Carnegie Mellon University
tom.mitchell@cmu.edu

CP16

Less Is More: Building Selective Anomaly Ensembles with Application to Event Detection in Temporal Graphs

Ensemble techniques for classification and clustering have long proven effective, yet anomaly ensembles have been barely studied. In this work, we tap into this gap and propose SELECT; an new ensemble approach for anomaly mining that employs novel techniques to automatically and systematically select the results to assemble in a fully unsupervised fashion. We apply our method to event detection in temporal graphs, where SELECT successfully utilizes five base detectors and seven consensus methods under a unified ensemble framework. We provide extensive quantitative evaluation of our approach on five real-world datasets. Thanks to its selection mechanism, SELECT yields superior performance compared to individual detectors alone, the full ensemble (naively combining all results), and an existing diversity-based ensemble.

Leman Akoglu
Stonybrook University
leman@cs.stonybrook.edu

Shebuti Rayana
Stony Brook University
srayana@cs.stonybrook.edu

PP1

Towards Classification of Social Streams

Social streams have become very popular in recent years because of the increasing popularity of social media sites such as *Twitter*, and *Facebook*. In this paper, we will focus on the classification problem for social streams. Unfortunately, such streams are extremely noisy, and contain large volumes of information, with information about *network linkages* between the participants exchanging messages. This is additional social information, associated with the text stream, which can be very helpful for classification. We combine an LSH method with an incremental SVM model in order to design an effective and efficient social context-sensitive streaming classifier for this scenario. The LSH model is used for learning the social context, and the SVM model is used for more effective classification within this context. We will present experimental results, which show the effectiveness of our techniques over a wide variety of other methods.

Charu C. Aggarwal
IBM T. J. Watson Research Center
charu@us.ibm.com

Min-Hsuan Last Tsai
Google
mhtsai@google.com

Thomas Huang
University of Illinois at Urbana-Champaign
huang@ifp.uiuc.edu

PP1

Mobile App Security Risk Assessment: A Crowdsourcing Ranking Approach from User Comments

Along with the exponential growth on markets of mobile Applications (apps), comes the serious public concern about the security and privacy issues. Therefore automatic app risk assessment becomes increasingly important to support users with useful evidences for their decisions. User comment provides a unique perspective from actual user experience, and should be considered valuable information source for risk assessment for mobile apps. In this paper, we provide a novel perspective to view the risk assessment of an app from its user comments as a crowdsourcing problem and adopt ranking model as the evaluation method. We develop a joint scheme to amalgamate feature learning and learning to rank models. Experiments conducted on two different real-world datasets show substantial performance improvements (*i.e., 6%–7%*) over the state-of-the-art methods.

Lei Cen
Purdue University
Purdue University
lcen@purdue.edu

Deguang Kong
University of Texas, Arlington
doogkong@gmail.com

Hongxia Jin
Samsung Research America
hongxia.jin@sisa.samsung.com

Luo Si
Purdue University
lsi@purdue.edu

PP1

Learning Stroke Treatment Progression Models for An MDP Clinical Decision Support System

We present a method for optimizing treatment plans for large, partially-observable, temporal clinical decision problems where the system is characterized only by a limited dataset of trials, such as those in EHR data. Model selection is applied to learn probabilistic temporal models directly from data, and approximate Markov Decision Process techniques are used to optimize policies according to a utility function. The method demonstrates promise for clinical decision support on the *International Stroke Trial* dataset.

Dan C. Coroian
Indiana University
dcoroian@indiana.edu

Kris Hauser
Duke University
kris.hauser@duke.edu

PP1

OnlineCM: Real-Time Consensus Classification with Missing Values

Combining predictions from multiple sources or models has been shown to be a useful technique in data mining. Unfortunately, as data are generated at an increasingly high speed, existing prediction aggregation methods are facing new challenges. The high velocity and volume of the data render existing batch mode prediction infeasible. And predictions from multiple models or data sources might not be perfectly synchronized, leading to abundant missing values in the prediction stream. We propose OnlineCM, to address the above challenges. OnlineCM keeps only a minimal yet sufficient footprint for both consensus prediction and missing value imputation over the prediction stream. Experiments demonstrates that OnlineCM achieves aggregated predictions that has close performance to the batch mode consensus maximization algorithm, and outperforms baseline methods significantly in large real world datasets.

Bowen Dong, Sihong Xie
University of Illinois at Chicago
bdong5@uic.edu, sxie6@uic.edu

Jing Gao
University at Buffalo
jing@buffalo.edu

Wei Fan
IBM T.J.Watson Research
wei.fan@gmail.com

Philip Yu
University of Illinois at Chicago
psyu@uic.edu

PP1

What shall I share and with Whom? - A Multi-Task Learning Formulation using Multi-Faceted Task Relationships

In multi-task learning, each task answers the question "Which other tasks should I share with"? A task may be differentially related to other tasks based on different feature subsets. Existing methods learn a single-faceted task relationship averaged over all features forcing a task to become similar to other tasks even on their unrelated features. We propose a novel multi-task learning model that learns multi-faceted task relationship, allowing tasks to collaborate differentially on different feature subsets.

Sunil K. Gupta
Deakin University, Geelong Waurn Ponds Campus
Victoria, Australia
sunil.gupta@deakin.edu.au

PP1

A Generalized Mixture Framework for Multi-Label Classification

We develop a novel probabilistic ensemble framework for multi-label classification based on the *mixtures-of-experts* architecture. We combine multi-label classification models in the *classifier chains family* that decompose the class posterior distribution $P(Y_1, ..., Y_d|\mathbf{X})$ using a product of posterior distributions over components of the output space. Accordingly, we recover a rich set of dependency relations among inputs and outputs that a single model cannot capture.

Charmgil Hong
Department of Computer Science
University of Pittsburgh
charmgil@cs.pitt.edu

Iyad Batal
GE Global Research
iyad.batal@ge.com

Milos Hauskrecht
University of Pittsburgh
milos@cs.pitt.edu

PP1

Domain-Knowledge Driven Cognitive Degradation Modeling for Alzheimers Disease

Cognitive monitoring and screening holds great promise for early detection of AD. A critical enabler is the personalized degradation model to predict the cognitive status over time. However, estimating such a model using individuals data faces challenges due to the sparsity and fragmented nature of the cognitive data of each individual. To mitigate this problem, we propose novel methods, called the collaborative degradation model (CDM) together with its extended network regularized version, the NCDM, which can incorporate useful domain knowledge into the degradation modeling. While NCDM results in a difficult optimization problem, we develop efficient algorithm to solve this problem and further provide theoretical results that ensure that the proposed algorithm can guarantee non-decreasing property. Both simulation studies and real-world application are conducted across different degradation models and sampling schemes, which demonstrate the superiority of the proposed methods over existing methods.

Shuai Huang
university of washington

University of Washington
shuaih@uw.edu

PP1

Optimizing Hashing Functions for Similarity Indexing in Arbitrary Metric and Nonmetric Spaces

A large number of methods have been proposed for similarity indexing in Euclidean spaces, and several such methods can also be used in arbitrary metric spaces. Such methods exploit specific properties of Euclidean spaces or general metric spaces. Designing general-purpose similarity indexing methods for arbitrary metric and non-metric distance measures is a more difficult problem, due to the vast heterogeneity of such spaces and the lack of common properties that can be exploited. In this paper, we propose a generally applicable method for similarity-based indexing in arbitrary metric and nonmetric spaces, based on hashing. We build upon the technique of Distance-Based Hashing (DBH), which organizes database objects in multiple hash tables, so that two similar objects tend to fall in the same bucket in at least one of those hash tables. The main contribution is in showing how to optimize the hashing functions for accuracy and efficiency, using training data.

Pat Jangyodsuk
University of Texas at Arlington
pat.jangyodsuk@mavs.uta.edu

Panagiotis Papapetrou
Birkbeck, University of London
panagiotis@dsv.su.se

Vassilis Athitsos
University of Texas at Arlington
athitsos@uta.edu

PP1

Ensemble Learning Methods for Binary Classification with Multi-Modality Within the Classes

We consider binary classification problems where each of the two classes show multi-modal distribution in the feature space. Inspired by existing ensemble learning methods for multi-class classification, we develop ensemble learning methods for binary classification that make use of the bipartite nature of the positive and negative modes in the data. We demonstrate the effectiveness of the proposed ensemble learning methods over a synthetic dataset and a real-world application involving global lake monitoring.

Anuj Karpatne
University of Minnesota
anuj@cs.umn.edu

Ankush Khandelwal
University of Minnesota- Twin Cities
ankush@cs.umn.edu

Vipin Kumar
University of Minnesota
kumar@cs.umn.edu

PP1

A Framework for Simplifying Trip Data into Networks Via Coupled Matrix Factorization

Portable devices such as GPS-equipped smart phones and cameras are able to provide detailed spatio-temporal trip event data for each user. Such data can be aggregated over many users to provide large amounts of behavioral data of very fine granularity. Trying to simplify this data into meaningful higher-level insights is challenging for a variety of reasons. In this paper we study the problem of simplifying spatio-temporal trip data and summarizing them into an easily interpretable graph/network. We propose several constrained coupled nonnegative matrix factorization formulations that simultaneously cluster locations and times based on the associated trips, and develop a (block) coordinate descent algorithm to solve them. We empirically evaluate our approach on a real world data set of taxis' GPS traces and show the advantages of our approach over traditional clustering algorithms.

Chia-Tung Kuo
Computer Science Department, University of California, Davis
tomkuo@ucdavis.edu

James Bailey
The University of Melbourne
baileyj@unimelb.edu.au

Ian Davidson
University of California, Davis
davidson@cs.ucdavis.edu

PP1

MET: A Fast Algorithm for Minimizing Propagation in Large Graphs with Small Eigen-Gaps

Given a graph G and a budget k, how can we quickly find the best k edges to delete that minimize dissemination in G? Stopping dissemination in a graph is important in a variety of fields from epidemiology to cyber security. The spread of an entity on an arbitrary graph G depends on two properties: the topology of G and the characteristics of the entity. In many settings, we cannot manipulate the latter. That leaves us with modifying the former by removing nodes and/or edges. We know that the largest eigenvalue of G's adjacency matrix is a good indicator for its path capacity. Thus, methods that quickly reduce the largest eigenvalue of G often minimize dissemination on G. But, a problem arises when the differences between the largest eigenvalues of G are small. This problem (a.k.a. the *small eigen-gap problem*) occurs often in social graphs. We present a scalable algorithm, called *MET*, that efficiently and effectively solves the small eigen-gap problem.

Long Le
Rutgers University
longtle@cs.rutgers.edu

Tina Eliassi-Rad
Department of Computer Science
Rutgers University
eliassi@cs.rutgers.edu

Hanghang Tong
Arizona State University
hanghang.tong@asu.edu

PP1

Learning Compressive Sensing Models for Big

Spatio-Temporal Data

Compressive sensing (CS) reconstructs the compressed signals exactly when incoming data can be sparsely represented with a fixed number of components. However, a real-world signal cannot be represented with the fixed number of components. We present the first CS framework that handles signals without the fixed sparsity assumption, which allows an analytic derivation of total error in our spatio-temporal Low Complexity Sampling (LCS). LCS requires shorter compressed signals than existing CS frameworks.

Dongeun Lee, Jaesik Choi
Ulsan National Institute of Science and Technology
eundong@unist.ac.kr, jaesik@unist.ac.kr

PP1

Multi-View Low-Rank Analysis for Outlier Detection

Outlier detection is a fundamental problem in data mining. Unlike most existing methods that are designed for single-view data, we propose a Multi-view Low-Rank Analysis (MLRA) framework in this paper. First, it performs cross-view low-rank analysis for revealing the intrinsic structures of data. Second, it identifies outliers by estimating the outlier score for each sample. Experimental results on seven UCI datasets and the USPS-MNIST dataset demonstrate the effectiveness of our approach.

Sheng Li, Ming Shao, Yun Fu
Northeastern University
shengli@ece.neu.edu, mingshao@ece.neu.edu, yunfu@ece.neu.edu

PP1

Reafum: Representative Approximate Frequent Subgraph Mining

Noisy graph data and pattern variations are two thorny problems faced by mining frequent subgraphs. Traditional methods only generate patterns that have enough perfect matches in the graph database. As a result, a pattern may either remain undetected or be reported as multiple patterns if it manifests slightly different instances in different graphs. We investigate the problem of approximate frequent pattern mining, with a focus on finding non-redundant representative frequent patterns that summarize the frequent patterns allowing approximate matches.

Ruirui Li, Wei Wang
University of California Los Angeles
goldenbearray@gmail.com, weiwang@cs.ucla.edu

PP1

Dias: A Disassemble-Assemble Framework for Highly Sparse Text Clustering

We propose a DIssemble-ASsemble (DIAS) framework for text clustering. DIAS employs simple random feature sampling to disassemble high-dimensional text data and gains diverse structural knowledge. Then the multi-view knowledge is assembled by weighted Information-theoretic Consensus Clustering. Extensive experiments demonstrate the advantages of DIAS over other widely used methods. In addition, it is the natural suitability to distributed computing that makes DIAS become a promising candidate for big text clustering.

Hongfu Liu
Northeastern University
leohf1029@foxmail.com

Junjie Wu
Beihang University
wujj@buaa.edu.cm

Dacheng Tao
University of Technology, Sydney
dacheng.tao@uts.edu.au

Yuchao Zhang
Beijing Institute of System Engineering
dragonzyc@163.com

Yun Fu
Northeastern University
yunfu@ece.neu.edu

PP1

Optimal Event Sequence Sanitization

Frequent event mining is a fundamental task to extract insight from an event sequence. However, it may expose sensitive events that leak confidential business knowledge or lead to intrusive inferences about groups of individuals. In this work, we aim to prevent this threat, by deleting occurrences of sensitive events, while preserving the utility of the event sequence. To quantify utility, we propose a model that captures changes, caused by deletion, to the probability distribution of events across the sequence. Based on the model, we define the problem of sanitizing an event sequence as an optimization problem. Solving the problem is important to preserve the output of many mining tasks, but also challenging. To optimally solve the problem when there is one sensitive event, we develop an efficient algorithm based on dynamic programming. The algorithm also forms the basis of a method that optimally sanitizes an event sequence, when there are multiple sensitive events.

Grigorios Loukides
Cardiff University
g.loukides@cs.cf.ac.uk

Robert Gwadera
EPFL
robert.gwadera@epfl.ch

PP1

Predicting Neighbor Distribution in Heterogeneous Information Networks

Recently, considerable attention has been devoted to the prediction problems arising from heterogeneous information networks. In this paper, we present a new prediction task, Neighbor Distribution Prediction (NDP), which aims at predicting the distribution of the labels on neighbors of a given node and is valuable for many different applications in heterogeneous information networks. We propose an Evolution Factor Model (EFM) for NDP, which utilizes two new structures proposed, i.e. Neighbor Distribution Vector (NDV) to represent the state of a given node's neighbors, and Neighbor Label Evolution Matrix (NLEM) to capture the dynamics of a neighbor distribution. To address data sparsity, we first cluster all nodes and learns an NLEM for

each cluster instead of for each node. For fairly evaluating, we propose a new metric: Virtual Accuracy (VA). Extensive experiments validate the effectiveness of our proposed model EFM and metric VA.

Yuchi Ma
Sichuan University
Chengdu, China
scu.Richard.Ma@gmail.com

Ning Yang
School of Computer Science
Sichuan University, Chengdu, China
ningyang@scu.edu.cn

Chuan Li, Lei Zhang
Sichuan University
Chengdu, China
lcharles@scu.edu.cn, leizhang@scu.edu.cn

Philip S. Yu
University of Illinois at Chicago
Chicago, USA
psyu@uic.edu

PP1

Temporally Coherent CRP: A Bayesian Non-Parametric Approach for Clustering Tracklets with Applications to Person Discovery in Videos

A video can be represented as a sequence of tracklets, each spanning over 10-20 successive frames, and each tracklet is associated with one entity (eg. person in case of TV-serial videos). Tracklets exhibit rich spatio-temporal structure. The task of Person Discovery in long TV-series videos can be naturally posed as tracklet clustering, and existing approaches give unsatisfactory performance on it. In this paper we attempt to leverage Temporal Coherence (TC) of videos to improve tracklet clustering through a Bayesian nonparametric approach: Temporally Coherent Chinese Restaurant Process (TC-CRP). On the task of discovering persons in TV serials via tracklet clustering, without meta-data such as scripts, TC-CRP shows up significant improvement compared to state-of-the-art parametric models. Moreover, unlike existing approaches TC-CRP can perform online tracklet clustering on streaming videos with very little performance deterioration, and can also automatically reject outliers.

Adway Mitra
Indian Institute of Science
adway.cse@gmail.com

Soma Biswas
Indian Institute of Science
Electrical Engineering
soma.biswas@ee.iisc.ernet.in

Chiranjib Bhattacharyya
Indian Institute of Science, Bangalore
India - 560012
chiru@csa.iisc.ernet.in

PP1

Correlating Surgical Vital Sign Quality with 30-Day Outcomes Using Regression on Time Series Segment Features

We present a method of feature extraction on surgical vital sign time series in conjunction with a regression based ordinal classifier to assign a quality label to each case based on expert assessment. We compare this approach to standard approaches to time series classification and use the best model to classify over 90,000 cases and correlate the labels with significant negative outcomes.

Risa Myers
Rice University
rbm2@rice.edu

John Frenzel, Joseph Ruiz
University of Texas MD Anderson Cancer Center
jfrenzel@mdanderson.org, jrruiz@mdanderson.org

Christopher Jermaine
Rice University
cmj4@rice.edu

PP1

Multi-Layered Framework for Modeling Relationships Between Biased Objects

The *Infinite Relational Model* (IRM) is a well-known relational model for discovering co-cluster structures with an unknown number of clusters. The IRM and several related models commonly assume that link probability between two objects depends only on their cluster assignment. However, relational models based on this assumption often lead us to extract many non-informative and unexpected clusters. This is because the cluster structures underlying real-world relationships are often blurred by *biases* that are inherent to individual objects. To overcome this problem, we propose a multi-layered framework that extracts a clear co-cluster structure in the presence of objects' biases. Then, we propose a new model which is a special instance of the proposed framework that incorporates the IRM. Experiments conducted using real-world datasets confirm that the proposed model successfully extracts clear and interpretable cluster structures from blurred relational data.

Iku Ohama
R&D Division, Panasonic Co., Ltd.
ohama.iku@jp.panasonic.com

Takuya Kida
Hokkaido University
kida@ist.hokudai.ac.jp

Hiroki Arimura
Graduate School of IST, Hokkaido University
arim@ist.hokudai.ac.jp

PP1

Specldа: Modeling Product Reviews and Specifications to Generate Augmented Specifications

Product specifications are often available for a product on E-commerce websites. However, novice customers often do not have enough knowledge to understand all features of a product, especially advanced features. In order to provide useful knowledge to the customers, we propose to automatically generate augmented product specifications, which contains relevant opinions for product feature values, feature importance, and product-specific words. Specifi-

cally, we propose a novel Specification Latent Dirichlet Allocation (SpecLDA) that can enable us to effectively model product reviews and specifications at the same time. Experiment results show that SpecLDA can effectively model product reviews with specifications. The model can be used for any text collections with specification (key-value) type prior knowledge.

Dae Hoon Park
University of Illinois at Urbana-Champaign
dpark34@illinois.edu

ChengXiang Zhai
University of Illinois at Urbana Champaign
czhai@illinois.edu

Lifan Guo
TCL Research America
guolifan@tcl.com

PP1

Mining Multi-Relational Gradual Patterns

Gradual patterns highlight covariations of attributes of the form *"The more/less X, the more/less Y"*. This paper extends the notion of gradual pattern to the case in which the co-variations are possibly expressed between attributes of different database relations. The interestingness measure for this class of "relational gradual patterns" is defined on the basis of both Kendall's τ and gradual supports. Moreover, this paper proposes two algorithms, named *τRGP Miner* and *gRGP Miner*, for the discovery of relational gradual rules. Three pruning strategies to reduce the search space are proposed. The efficiency of the algorithms is empirically validated, and the usefulness of relational gradual patterns is proved on some real-world databases.

Nhathai Phan
CIS Department, University of Oregon
haiphan@cs.uoregon.edu

PP1

Modeling User Arguments, Interactions, and Attributes for Stance Prediction in Online Debate Forums

Online debate forums are important social media for people to voice their opinions and debate with each other. Mining user stances or viewpoints from these forums has been a popular research topic. However, most current work does not address an important problem: for a specific issue, there may not be many users participating and expressing their opinions. Despite the sparsity of user stances, users may provide rich side information; for example, users may write arguments to back up their stances, interact with each other, and provide biographical information. In this work, we propose an integrated model to leverage side information. Our proposed method is a regression-based latent factor model which jointly models user arguments, interactions, and attributes. Our method can perform stance prediction for both warm-start and cold-start users. We demonstrate in experiments that our method has promising results on both micro-level and macro-level stance prediction.

Minghui Qiu
Singapore Management University
minghui.qiu.2010@smu.edu.sg

Yanchuan Sim, Noah Smith
Carnegie Mellon University
ysim@cs.cmu.edu, nasmith@cs.cmu.edu

Jing Jiang
Singapore Management University
jingjiang@smu.edu.sg

PP1

Predicting Preference Tags to Improve Item Recommendation

Collaborative filtering (CF) based recommender systems identify and recommend interesting items to a given user based on the users past rating activity. These systems improve their recommendations by identifying user preferences and item related information from external sources, like reviews written by users, or concept tags shared by users about these items. These preferences are often reflected through a multi-criterion rating. In this study, we seek to improve recommender systems by integrating user preferences as side information within standard neighborhood-based and matrix factorization based methods.

Huzefa Rangwala, Tanwistha Saha, Carlotta Domeniconi
George Mason University
rangwala@cs.gmu.edu, tsaha@masonlive.gmu.edu, carlotta@cs.gmu.edu

PP1

Data Stream Classification Guided by Clustering on Nonstationary Environments and Extreme Verification Latency

Data stream classification algorithms for nonstationary environments frequently assume the availability of labels after the classification. However, many applications involve high costs to obtain these labels. Such a scenario in which the actual labels of processed data are never available is called extreme verification latency. We present an algorithm to classify nonstationary data in this scenario. Our method (SCARGC) consists of a clustering followed by a classification step applied repeatedly in a closed loop fashion.

Vinicius Souza, Diego Silva
University of Sao Paulo
vsouza@icmc.usp.br, diegofsilva@gmail.com

Joao Gama
University of Porto
joaojgama@gmail.com

Gustavo Batista
University of Sao Paulo
hustav@gmail.com

PP1

Mining Block I/O Traces for Cache Preloading with Sparse Temporal Non-Parametric Mixture of Multivariate Poisson

Existing caching strategies, though well suited to exploit short range spatio-temporal patterns, are unable to leverage long-range motifs for improving hitrates. Motivated by this, we investigate novel Bayesian non-parametric modeling(BNP) techniques for count vectors, to capture long range correlations for cache preloading, by mining Block

I/O traces. We propose a DP based mixture model of Multivariate Poisson and its temporal extension for non-parametric clustering of count vectors. However the Multivariate Piosson is computationally expensive for high dimensional data. Hence, we exploit sparsity in our data and introduce the Sparse DP mixture of multivariate Poisson, and its temporal extension, leading to efficient inference. Finally, we propose an algorithm for cache preloading using our models taking the first step towards capturing long range patterns in storage traces for cache preloading. Experimentally, we show a dramatic improvement in hitrates on benchmark traces.

Lavanya S. Tekumalla
Dept of CSA, Indian Institute Of Science
lavanya.iisc@gmail.com

Chiranjib Bhattacharyya
Indian Institute of Science, Bangalore
India - 560012
chiru@csa.iisc.ernet.in

PP1

Taming the Empirical Hubness Risk in Many Dimensions

The hubness phenomenon has recently come into focus as an important aspect of the curse of dimensionality that affects many instance-based learning systems. In this talk, we introduce the concept of relative hubness risk and re-evaluate several recently proposed metric learning approaches that propose to reduce the overall hubness in the data in order to ensure better and more robust system performance.

Nenad Tomaev
Artificial Intelligence Laboratory, Jozef Stefan Institute
nenad.tomasev@gmail.com

PP1

Scalable Clustering of Time Series with U-Shapelets

A recently introduced primitive for time series data mining, unsupervised shapelets (u-shapelets), demonstrated significant potential for time series clustering. They have several advantages over rival methods: find and consider only (potentially) relevant subsequences to clustering; defined when the time series are of different lengths; mitigate sensitivity to irrelevant data (e.g., noise, dropouts); provide insights into the data. Unfortunately, state-of-the-art algorithms for u-shapelets search are intractable and so their advantages only demonstrated on tiny datasets. We propose a simple approach to speed up a u-shapelet discovery by two orders of magnitude, without significant loss in clustering quality.

Liudmila Ulanova
University of California, Riverside
lulan001@ucr.edu

PP1

Causal Inference by Direction of Information

We propose a new principle for causal inference based on Kolmogorov complexity. In a nutshell, we determine how one data object helps to describe the other, to identify the most likely causal direction by the strongest *direction* of information. To put this to practice, we propose Ergo, a non-parametric instantiation for multivariate real-valued data. Empirical evaluation shows that Ergo is robust against noise and dimensionality, efficient, and outperforms the state of the art by a wide margin.

Jilles Vreeken
Max Planck Institute for Informatics
Saarland University
jilles@mpi-inf.mpg.de

PP1

Graph Regularized Meta-path Based Transductive Regression in Heterogeneous Information Network

A number of real-world networks are heterogeneous information networks, which are composed of different types of nodes and links. Numerical prediction in heterogeneous information networks is a challenging but significant area because network based information for unlabeled objects is usually limited to make precise estimations. In this paper, we consider a graph regularized meta-path based transductive regression model (Grempt), which combines the principal philosophies of typical graph-based transductive classification methods and transductive regression models designed for homogeneous networks. The computation of our method is time and space efficient and the precision of our model can be verified by numerical experiments.

Mengting Wan, Yunbo Ouyang
University of Illinois, Urbana-Champaign
mwan5@illinois.edu, youyang4@illinois.edu

Lance Kaplan
U.S. Army Research Laboratory
lance.m.kaplan@us.army.mil

Jiawei Han
UIUC
hanj@illinois.edu

PP1

Localizing Temporal Anomalies in Large Evolving Graphs

Mining for anomalies in graph structured datasets is an important and challenging problem for many applications including security, health care, and social media. In this paper, we propose a novel framework to localize temporal anomalies in large evolving graphs with reduced false alarm rate. Specifically, we first introduce a node-centric model based on Vector Autoregression to analyze node behavior history in dynamic graphs. Then we develop two community-centric models to reduce the amount of false positive results by tracking the structural change and dynamics of graph communities. We analyze the performance of our proposed anomaly localization framework on several synthetic and real-world data sets including Enron email network data, an enterprise network traffic data, and CNN public Facebook page. All experimental results show the effectiveness and consistency of our framework in localizing temporal anomalies with reduced false alarm rate.

Teng Wang
University of California, Davis
tewang@ucdavis.edu

CHUNSHENG Fang, DEREK Lin
Pivotal Software, Inc.

cfang@pivotal.io, dlin@pivotal.io

S.FELIX Wu
University of California, Davis
sfwu@ucdavis.edu

PP1

Non-Exhaustive, Overlapping k-Means

In real datasets, clusters can overlap and there are often outliers that do not belong to any cluster. We propose an intuitive objective function that captures the issues of overlap and non-exhaustiveness in a unified manner. To optimize this objective, we develop a simple iterative algorithm which we call NEO-K-Means (Non-Exhaustive Overlapping K-Means). Furthermore, by considering an extension to weighted kernel k-means, we can apply our NEO-K-Means algorithm to the overlapping community detection problem.

Joyce J. Whang
The University of Texas at Austin
joyce@cs.utexas.edu

Inderjit S. Dhillon
University of Texas at Austin
inderjit@cs.utexas.edu

David F. Gleich
Purdue University
dgleich@purdue.edu

PP1

Festival, Date and Limit Line: Predicting Vehicle Accident Rate in Beijing

Thousands of vehicle accidents happen every day in Beijing, leading to huge losses. Government traffic management bureau, hospitals, and insurance companies put massive manpower and material resources to deal with accidents. For more reasonable resource assignment, in this study we focus on the prediction of daily *Vehicle Accident Rate* (*VAR*), namely the percentage of vehicles with accidents. Specifically, we analyze how the variation of *VAR* correlates with the macroscopic features, like Chinese festival, date, tail-number limit line etc., and develop the prediction model for *VAR* based on these features. Our analysis is based on the records of two-year accidents on the vehicles, which are insured by a local insurance giant in Beijing. Experiments show that the proposed model can predict the long-term *VAR* for at least three months in advance, with satisfactory results. Note also that our study is based on the local conditions in Beijing with Chinese characteristics. It not only helps government bureaus and insurance companies to operate more efficiently, but also helps to know many underlying characteristics of this China capital in a macroscopic perspective.

Xinyu Wu
ICT, CAS
UCAS
graduatewuxinyu@126.com

Ping Luo
ICT, CAS
luop@ict.ac.cn

Qing He
Room 506, Kexueyuan Nanlu #6, Zhongguan Cun,
Haidian District, Beijing, China
heq@ics.ict.ac.cn

Tianshu Feng
USTC
tsfeng@mail.ustc.edu.cn

Fuzhen Zhuang
ICT, CAS
zhuangfz@ics.ict.ac.cn

PP1

A Multi-Label Least-Squares Hashing For Scalable Image Search

We propose a Multi-label Least-Squares Hashing (MLSH) method for multi-label data hashing. It can directly work well on multi-label data. MLSH first utilizes the equivalent form of CCA and Least-Squares to project original multi-label data into lower-dimensional space; then, in the lower-dimensional space, it learns the project matrix and gets final binary codes of data. Experimental results show that MLSH outperforms several state-of-the-art hashing methods.

Xin-Shun Xu, Shengsheng Wang
Shandong University
xuxinshun@sdu.edu.cn, xiaoshengforever@foxmail.com

Zi Huang
The University of Queensland
huang@itee.uq.edu.au

PP1

Simpleppt: A Simple Principal Tree Algorithm

We develop a new model, which captures the local information of the underlying graph structure based on reversed graph embedding. As a special case, a principal tree model is proposed and a new algorithm is developed that learns a tree structure automatically from data. The new algorithm is simple and parameter-free with guaranteed convergence. Experimental show that the proposed method compares favorably with baselines and can discover a breast cancer progression path with multiple branches.

Qi Mao
The State University of New York at Buffalo
maoq1984@gmail.com

Le Yang
Department of Computer Science and Engineering
The State University of New York at Buffalo
lyang25@buffalo.edu

Li Wang
The Institute for Computational and Experimental
Research in
Brown University
liwangucsd@gmail.com

Steve Goodison
Department of Health Sciences Researc
Mayo Clinic
goodisonsteve@gmail.com

Yijun Sun

Department of Microbiology and Immunology
The State University of New York at Buffalo
yijunsun@buffalo.edu

PP1
Spatiotemporal Event Forecasting in Social Media

Event forecasting in Twitter is an important and challenging problem. Most existing approaches focus on forecasting temporal events (such as elections and sports) and do not consider spatial features and their underlying correlations. In this paper, we propose a generative model for spatiotemporal event forecasting in Twitter. Our model characterizes the underlying development of future events by jointly modeling the structural contexts and spatiotemporal burstiness. An effective inference algorithm is developed to train the model parameters. Utilizing the trained model, the alignment likelihood of tweet sequences is calculated by dynamic programming. Extensive experimental evaluations on two different domains demonstrated the effectiveness of our proposed approach.

Liang Zhao
Virginia Tech
liangz8@vt.edu

Feng Chen
University of Albany-SUNY
fchen5@albany.edu

Chang-Tien Lu
Computer Science, Virginia Tech
ctlu@vt.edu

Naren Ramakrishnan
Computer Science
Virginia Tech
naren@cs.vt.edu