

SIAM International Conference on Data Mining 2015 (SDM15)

Vancouver, Canada
30 April – 2 May 2015

Editors:

**Suresh Venkatasubramanian
Jieping Ye**

ISBN: 978-1-5108-1152-2

Printed from e-media with permission by:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571



Some format issues inherent in the e-media version may also appear in this print version.

Copyright© (2015) by SIAM: Society for Industrial and Applied Mathematics
All rights reserved.

Printed by Curran Associates, Inc. (2015)

For permission requests, please contact SIAM: Society for Industrial and Applied Mathematics
at the address below.

SIAM
3600 Market Street, 6th Floor
Philadelphia, PA 19104-2688 USA

Phone: (215) 382-9800
Fax: (215) 386-7999

siambooks@siam.org

Additional copies of this publication are available from:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: 845-758-0400
Fax: 845-758-2634
Email: curran@proceedings.com
Web: www.proceedings.com

Table of Contents

Session 1: Networks, Graphs I

Functional Node Detection on Linked Data	1
Kang Li, Jing Gao, Suxin Guo, Nan Du, Aidong Zhang	
Where Graph Topology Matters: The Robust Subgraph Problem.....	10
Hau Chan, Shuchu Han, Leman Akoglu	
Same bang, fewer bucks: Efficient discovery of the cost-influence skyline	19
Matthijs van Leeuwen, Antti Ukkonen	
Selecting shortcuts for a smaller world	28
Nikos Parotsidis, Evaggelia Pitoura, Panayiotis Tsaparas	
Significant Subgraph Mining with Multiple Testing Correction	37
Mahito Sugiyama, Felipe Llinares López, Niklas Kasenburg, Karsten Borgwardt	

Session 2: Metric Learning, Feature Selection/Extraction

From Categorical to Numerical: Multiple Transitive Distance Learning and Embedding	46
Kai Zhang, Qiaojun Wang, Zhengzhang Chen, Ivan Marsic, Vipin Kumar, Guofei Jiang, Jie Zhang	
Spectral Embedding of Signed Networks	55
Quan Zheng, David Skillicorn	
An LLE based Heterogeneous Metric Learning for Cross-media Retrieval.....	64
Peng Zhou, Liang Du, Mingyu Fan, Yi-Dong Shen	
Feature Selection for Nonlinear Regression and its Application to Cancer Research	73
Yijun Sun, Jin Yao, Steve Goodison	
Efficient Partial Order-preserving Unsupervised Feature Selection on Networks	82
Xiaokai Wei, Sihong Xie, Philip S. Yu	

Session 3: Clustering

NetCodec: Community Detection from Individual Activities.....	91
Long Tran, Mehrdad Farajtabar, Le Song, Hongyuan Zha	

Efficient Algorithms for a Robust Modularity-Driven Clustering of Attributed Graphs	100
Patricia Iglesias Sanchez, Emmanuel Müller, Uwe Leo Korn, Klemens Böhm, Andrea Kappes, Tanja Hartmann, Dorothea Wagner	
Vertex Clustering of Augmented Graph Streams	109
Ryan McConville, Weiru Liu, Paul Miller	
Tensor Spectral Clustering for Partitioning Higher-order Network Structures	118
Austin Benson, David Gleich, Jure Leskovec	
Community Detection for Emerging Networks.....	127
Jiawei Zhang, Philip S. Yu	

Session 4: Applications

Labeling Educational Content with Academic Learning Standards	136
Danish Contractor, Kashyap Popat, Shajith Ikbal, Sumit Negi, Bikram Sengupta, Mukesh Mohania	
Data mining for real mining: A robust algorithm for prospectivity mapping with uncertainties.....	145
Justin Granek, Eldad Haber	
Product Adoption Rate Prediction: A Multi-factor View	154
Le Wu, Qi Liu, Enhong Chen, Xing Xie, Chang Tan	
PatentCom: A Comparative View of Patent Document Retrieval	163
Longhui Zhang, Lei Li, Chao Shen, Tao Li	
Combating Product Review Spam Campaigns via Multiple Heterogeneous Pairwise Features	172
Chang Xu, Jie Zhang	

Session 5: Recommendation, Classification

A Bayesian Framework for Modeling Human Evaluations.....	181
Himabindu Lakkaraju, Jure Leskovec, Jon Kleinberg, Sendhil Mullainathan	
Feature-based factorized Bilinear Similarity Model for Cold-Start Top-n Item Recommendation.....	190
Mohit Sharma, Jiayu Zhou, Junling Hu, George Karypis	
Cross-Modal Retrieval: A Pairwise Classification Approach.....	199
Aditya Menon, Didi Surian, Sanjay Chawla	

Binary classifier calibration using a Bayesian non-parametric approach..... 208
Mahdi Pakdaman Naeini, Gregory Cooper, Milos Hauskrecht

Semi-supervised learning for structured regression on partially observed attributed graphs 217
Jelena Stojanovic, Milos Jovanovic, Djordje Gligorijevic, Zoran Obradovic

Session 6: Security/Privacy, Social Media

Health Insurance Market Risk Assessment: Covariate Shift and k-Anonymity 226
Dennis Wei, Karthikeyan Natesan Ramamurthy, Kush Varshney

Attacking DBSCAN for Fun and Profit..... 235
Jonathan Crussell, Philip Kegelmeyer

Result Integrity Verification of Outsourced Privacy-preserving Frequent Itemset Mining 244
Ruilin Liu, Wendy Wang

Modeling Users' Adoption Behaviors with Social Selection and Influence 253
Ziqi Liu, Fei Wang, Qinghua Zheng

Exploring the Impact of Dynamic Mutual Influence on Social Event Participation 262
Tong Xu, Hao Zhong, Hengshu Zhu, Hui Xiong, Enhong Chen, Guannan Liu

Session 7: Time Series, Online Learning

Efficient Online Relative Comparison Kernel Learning..... 271
Eric Heim, Matthew Berger, Lee Seversky, Milos Hauskrecht

Cheetah: Fast Graph Kernel Tracking on Dynamic Graphs..... 280
Liangyue Li, Hanghang Tong, Yanghua Xiao, Wei Fan

On the Non-Trivial Generalization of Dynamic Time Warping to the Multi-Dimensional Case 289
Mohammad Shokoohi-Yekta, Jun Wang, Eamonn Keogh

Fast Mining of a Network of Coevolving Time Series 298
Yongjie Cai, Hanghang Tong, Wei Fan, Ping Ji

Shapelet Ensemble for Multi-dimensional Time Series..... 307
Mustafa S Cetin, Abdullah Mueen, Vince D. Calhoun

Session 8: Matrix/Tensor

Low Rank Representation on Riemannian Manifold of Symmetric Positive Definite Matrices	316
Yifan Fu, Junbin Gao, Xia Hong, David Tien	
Getting to Know the Unknown Unknowns: Destructive-Noise Resistant Boolean Matrix Factorization	325
Sanjar Karaev, Pauli Miettinen, Jilles Vreeken	
Convex Matrix Completion: A Trace-Ball Optimization Perspective	334
Guangxiang Zeng, Ping Luo, Enhong Chen, Hui Xiong, Hengshu Zhu, Qi Liu	
Near-separable Non-negative Matrix Factorization with ℓ_1- and Bregman Loss Functions	343
Abhishek Kumar, Vikas Sindhwani	
Personalized TV Recommendation with Mixture Probabilistic Matrix Factorization	352
Huayu Li, Hengshu Zhu, Yong Ge, Yanjie Fu, Yuan Ge	

Session 9: Multi-source and Heterogeneous Learning

Legislative Prediction with Dual Uncertainty Minimization from Heterogeneous Information	361
Yu Cheng, Ankit Agrawal, Huan Liu, Alok Choudhary	
PLUMS: Predicting Links Using Multiple Sources	370
Karthik Subbian, Arindam Banerjee, Sugato Basu	
SourceSeer: Forecasting Rare Disease Outbreaks Using Multiple Data Sources	379
Theodoros Rekatsinas, Saurav Ghosh, Sumiko Mekar, Elaine Nsoesie, John Brownstein, Lise Getoor, Naren Ramakrishnan	
GIN: A Clustering Model for Capturing Dual Heterogeneity in Networked Data	388
Jialu Liu, Chi Wang, Jing Gao, Quanquan Gu, Charu Aggarwal, Lance Kaplan, Jiawei Han	
Believe It Today or Tomorrow? Detecting Untrustworthy Information from Dynamic Multi-Source Data	397
Houping Xiao, Yaliang Li, Jing Gao, Fei Wang, Liang Ge, Wei Fan, Long Vu, Deepak Turaga	

Session 10: Networks, Graphs II

Rare Class Detection in Networks	406
Karthik Subbian, Charu C. Aggarwal, Jaideep Srivastava, Vipin Kumar	

Hidden Hazards: Finding Missing Nodes in Large Graph Epidemics.....	415
Shashidhar Sundareisan, Jilles Vreeken, B. Aditya Prakash	
Clustering and Ranking in Heterogeneous Information Networks via Gamma-Poisson Model.....	424
Junxiang Chen, Wei Dai, Yizhou Sun, Jennifer Dy	
A Divide-and-Conquer Algorithm for Betweenness Centrality.....	433
Dora Erdos, Vatche Ishakian, Azer Bestavros, Evimaria Terzi	
Frameworks to Encode User Preferences for Inferring Topic-sensitive Information Networks	442
Qingbo Hu, Sihong Xie, Shuyang Lin, Wei Fan, Philip Yu	

Session 11: Optimization

Dropout Training of Matrix Factorization and Autoencoders for Link Prediction in Sparse Graphs.....	451
Shuangfei Zhai, Zhongfei (Mark) Zhang	
An ADMM Algorithm for Clustering Partially Observed Networks.....	460
Necdet Serhat Aybat, Sahar Zarmehri, Soundar Kumara	
Scaling log-linear analysis to datasets with thousands of variables	469
Francois Petitjean, Geoffrey Webb	
A Distributed Frank-Wolfe Algorithm for Communication-Efficient Sparse Learning.....	478
Aurélien Bellet, Yingyu Liang, Alireza Bagheri Garakani, Maria-Florina Balcan, Fei Sha	
Exceptional Model Mining with Tree-Constrained Gradient Ascent	487
Thomas E. Krak, Ad Feelders	

Session 12: Multi-task/Transfer Learning

Formula: FactORized MUlti-task LeARning for task discovery in personalized medical models	496
Jianpeng Xu, Jiayu Zhou, Pang-Ning Tan	
Active Multi-task Learning via Bandits.....	505
Meng Fang, Dacheng Tao	
Hierarchical Active Transfer Learning	514
David Kale, Marjan Ghazvininejad, Anil Ramakrishna, Jingrui He, Yan Liu	
Learning Complex Rare Categories with Dual Heterogeneity	523
Pei Yang, Jingrui He, Jia-Yu Pan	

Faster Jobs in Distributed Data Processing using Multi-Task Learning..... 532
Neeraja Yadwadkar, Bharath Hariharan, Joseph Gonzalez, Randy Katz

Session 13: Text Mining, Applications Part I of II

**Selecting Social Media Responses to News: A Convex Framework Based
On Data Reconstruction 541**
Zaiyi Chen, Linli Xu, Enhong Chen, Zhefeng Wang, Biao Chang, Yitan Li

Tracking Events Using Time-dependent Hierarchical Dirichlet Tree Model 550
Rumeng Li, Tao Wang, Xun Wang

Session 14: Networks, Graphs, Applications Part I of II

Fast Eigen-Functions Tracking on Dynamic Graphs 559
Chen Chen, Hanghang Tong

**Approximation Algorithms for Reducing the Spectral Radius to Control
Epidemic Spread 568**
Sudip Saha, Abhijin Adiga, B. Aditya Prakash, Anil Vullikanti

Session 15: Text Mining, Applications Part II of II

Propagation-based Sentiment Analysis for Microblogging Data 577
Jiliang Tang, Chikashi Nobata, Anlei Dong, Yi Chang, Huan Liu

Polyglot-NER: Massive Multilingual Named Entity Recognition 586
Rami Al-Rfou, Vivek Kulkarni, Bryan Perozzi, Steven Skiena

Online Resource Allocation with Structured Diversification..... 595
Nicholas Johnson, Arindam Banerjee

**Towards Permission Request Prediction on Mobile Apps via Structure
Feature Learning 604**
Deguang Kong, Hongxia Jin

Session 16: Networks, Graphs, Applications Part II of II

On Influential Nodes Tracking in Dynamic Social Networks 613
Xiaodong Chen, Guojie Song, Xinran He, Kunqing Xie

**Less is More: Building Selective Anomaly Ensembles with Application
to Event Detection in Temporal Graphs 622**
Shebuti Rayana, Leman Akoglu

Principled Neuro-Functional Connectivity Discovery	631
Kejun Huang, Nicholas Sidiropoulos, Evangelos Papalexakis, Christos Faloutsos, Partha Talukdar, Tom Mitchell	
Estimating Ad Impact on Clicker Conversions for Causal Attribution: A Potential Outcomes Approach	640
Joel Barajas, Ram Akella, Aaron Flores, Marius Holtan	
 Posters	
Towards Classification of Social Streams	649
Min-Hsuan Tsai, Charu Aggarwal, Thomas Huang	
Mobile App Security Risk Assessment: A Crowdsourcing Ranking Approach from User Comments	658
Lei Cen, Deguang Kong, Hongxia Jin, Luo Si	
Learning Compressive Sensing Models for Big Spatio-Temporal Data	667
Dongeun Lee, Jaesik Choi	
Learning Stroke Treatment Progression Models for an MDP Clinical Decision Support System	676
Dan C. Coroian, Kris Hauser	
OnlineCM: Online Consensus Maximization with Missing Values.....	685
Bowen Dong, Sihong Xie, Jing Gao, Wei Fan, Philip S. Yu	
MET: A Fast Algorithm for Minimizing Propagation in Large Graphs with Small Eigen-Gaps	694
Long Le, Tina Eliassi-Rad, Hanghang Tong	
What shall I share and with Whom? - A Multi-Task Learning Formulation using Multi-Faceted Task Relationships.....	703
Sunil Gupta, Santu Rana, Dinh Phung, Svetha Venkatesh	
A Generalized Mixture Framework for Multi-label Classification.....	712
Charmgil Hong, Iyad Batal, Milos Hauskrecht	
Domain-Knowledge Driven Cognitive Degradation Modeling for Alzheimer’s Disease	721
Ying Lin, Kaibo Liu, Eunshin Byon, Xiaoning Qian, Shuai Huang	
Ensemble Learning Methods for Binary Classification with Multi-modality within the Classes	730
Anuj Karpatne, Ankush Khandelwal, Vipin Kumar	

A Framework for Simplifying Trip Data into Networks via Coupled Matrix Factorization.....	739
Chia-Tung Kuo, James Bailey, Ian Davidson	
Multi-view Low-Rank Analysis for Outlier Detection.....	748
Sheng Li, Ming Shao, Yun Fu	
REAFUM: Representative Approximate Frequent Subgraph Mining.....	757
Ruirui Li, Wei Wang	
DIAS: A Disassemble-Assemble Framework for Highly Sparse Text Clustering	766
Hongfu Liu, Junjie Wu, Dacheng Tao, Yuchao Zhang, Yun Fu	
Optimal event sequence sanitization	775
Grigorios Loukides, Robert Gwadera	
Predicting Neighbor Distribution in Heterogeneous Information Networks	784
Yuchi Ma, Ning Yang, Chuan Li, Lei Zhang, Philip S. Yu	
SimplePPT: A Simple Principal Tree Algorithm.....	792
Qi Mao, Le Yang, Li Wang, Steve Goodison, Yijun Sun	
Temporally Coherent CRP: A Bayesian Non-Parametric Approach for Clustering Tracklets with applications to Person Discovery in Videos	801
Adway Mitra, Soma Biswas, Chiranjib Bhattacharyya	
Correlating Surgical Vital Sign Quality with 30-Day Outcomes using Regression on Time Series Segment Features.....	810
Risa Myers, John Frenzel, Joseph Ruiz, Christopher Jermaine	
Multi-Layered Framework for Modeling Relationships between Biased Objects.....	819
Iku Ohama, Takuya Kida, Hiroki Arimura	
Optimizing Hashing Functions for Similarity Indexing in Arbitrary Metric and Nonmetric Spaces	828
Pat Jangyodsuk, Panagiotis Papapetrou, Vassilis Athitsos	
SpecLDA: Modeling Product Reviews and Specifications to Generate Augmented Specifications	837
Dae Hoon Park, ChengXiang Zhai, Lifan Guo	
Mining Multi-Relational Gradual Patterns.....	846
NhatHai Phan, Dino Ienco, Donato Malerba, Pascal Poncelet, Maguelonne Teisseire	
Modeling User Arguments, Interactions, and Attributes for Stance Prediction in Online Debate Forums	855
Minghui Qiu, Yanchuan Sim, Noah Smith, Jing Jiang	

Predicting Preference Tags to Improve Item Recommendation	864
Tanwistha Saha, Huzefa Rangwala, Carlotta Domeniconi	
Data Stream Classification Guided by Clustering on Nonstationary Environments and Extreme Verification Latency	873
Vinicius Souza, Diego Silva, Joao Gama, Gustavo Batista	
Mining Block I/O Traces for Cache Preloading with Sparse Temporal Non-parametric Mixture of Multivariate Poisson	882
Lavanya Sita Tekumalla, Chiranjib Bhattacharyya	
Taming the Empirical Hubness Risk in Many Dimensions	891
Nenad Tomašev	
Scalable Clustering of Time Series with U-Shapelets	900
Liudmila Ulanova, Nurjahan Begum, Eamonn Keogh	
Causal Inference by Direction of Information	909
Jilles Vreeken	
Graph Regularized Meta-path Based Transductive Regression in Heterogeneous Information Network.....	918
Mengting Wan, Yunbo Ouyang, Lance Kaplan, Jiawei Han	
Localizing Temporal Anomalies in Large Evolving Graphs.....	927
Teng Wang, Chunsheng Fang, Derek Lin, S. Felix Wu	
Non-exhaustive, Overlapping k-means	936
Joyce Whang, Inderjit Dhillon, David Gleich	
Festival, Date and Limit Line: Predicting Vehicle Accident Rate in Beijing	945
Xinyu Wu, Ping Luo, Qing He, Tianshu Feng, Fuzhen Zhuang	
A Multi-label Least-Squares Hashing For Scalable Image Search	954
Shengsheng Wang, Zi Huang, Xin-Shun Xu	
Spatiotemporal Event Forecasting in Social Media.....	963
Liang Zhao, Feng Chen, Chang-Tien Lu, Naren Ramakrishnan	