

12th SIAM International Conference on Data Mining 2012

**Anaheim, California, USA
26-28 April 2012**

Volume 1 of 2

ISBN: 978-1-62276-094-7

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© (2012) by SIAM: Society for Industrial and Applied Mathematics
All rights reserved.

Printed by Curran Associates, Inc. (2012)

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 CP1 - Applications – Climate and Geography

1 [Detecting and Tracking Coordinated Groups in Dense, Systematically Moving, Crowds](#)

James Rosswog and Kanad Ghose

12 [Large-Scale Nonparametric Estimation of Vehicle Travel Time Distributions](#)

Rikiya Takahashi, Takayuki Osogami and Tetsuro Morimura

24 [Drought Detection of the Last Century: An MRF-based Approach](#)

Qiang Fu, Arindam Banerjee, Stefan Liess and Peter Snyder

35 [Toward Data-driven, Semi-automatic Inference of Phenomenological Physical Models: Application to Eastern Sahel Rainfall](#)

Saurabh Pendse, Isaac Tetteh, Fredrick Semazzi, Vipin Kumar and Nagiza Samatova

47 [Sparse Group Lasso: Consistency and Climate Applications](#)

Soumyadeep Chatterjee, Karsten Steinhaeuser, Arindam Banerjee, Snigdhanu Chatterjee and Auroop Ganguly

Session CP2 - Clustering

59 [The Multi-Set Stream Clustering Problem](#)

Charu Aggarwal

70 [Stratification Based Hierarchical Clustering Over a Deep Web Data Source](#)

Tantan Liu and Gagan Agrawal

82 [Cluster-Aware Compression with Provable K-means Preservation](#)

Nikolaos Freris, Michail Vlachos and Deepak Turaga

94 [Supervised Clustering of Label Ranking Data](#)

Mihajlo Grbovic, Nemanja Djuric and Slobodan Vucetic

106 [Symmetric Nonnegative Matrix Factorization for Graph Clustering](#)

Da Kuang, Haesun Park and Chris Ding

Session - CP3 - Social Media

118 [Feature Selection with Linked Data in Social Media](#)

Jiliang Tang and Huan Liu

129 [Microscopic Social Influence](#)

Ting Wang, Mudhakar Srivatsa, Dakshi Agrawal and Ling Liu

141 [HAR: Hub, Authority and Relevance Scores in Multi-Relational Data for Query Search](#)

Xutao Li, Michael K. Ng and Yunming Ye

153 [Evaluating Event Credibility on Twitter](#)

Manish Gupta, Peixiang Zhao and Jiawei Han

165 [Multi-skill Collaborative Teams based on Densest Subgraphs](#)

Amita Gajewar and Atish Das Sarma

Session CP4 - Multi-Source and Multi-Task

177 [Heterogeneous Data Fusion via Space Alignment Using Nonmetric Multidimensional Scaling](#)

Jaegul Choo, Shawn Bohn, Grant Nakamura, Amanda White and Haesun Park

189 [Heterogeneous Datasets Representation and Learning using Diffusion Maps and Laplacian Pyramids](#)

Neta Rabin and Ronald Coifman

200 [A Bayesian Nonparametric Joint Factor Model for Learning Shared and Individual Subspaces from Multiple Data Sources](#)

Sunil Gupta, Dinh Phung and Svetha Venkatesh

212 [Adaptive Multi-task Sparse Learning with an Application to fMRI Study](#)

Xi Chen, Jingrui He, Rick Lawrence and Jaime Carbonell

224 [Learning from Heterogeneous Sources via Gradient Boosting Consensus](#)

Xiaoxiao Shi, JEAN-FRANCOIS PAIEMENT, David Grangier and Philip Yu

Session CP5 - Pattern Mining

236 [Slim: Directly Mining Descriptive Patterns](#)

Koen Smets and Jilles Vreeken

248 [MARBLES: Mining Association Rules Buried in Long Event Sequences](#)

Boris Cule, Nikolaj Tatti and Bart Goethals

260 [Mining Patterns in Networks using Homomorphism](#)

Anton Dries and Siegfried Nijssen

272 [Scalable Induction of Probabilistic Real-Time Automata Using Maximum Frequent Pattern Based Clustering](#)

Jana Schmidt, Sonja Ansorge and Stefan Kramer

284 [Class Relevant Pattern Mining in Output-Polynomial Time](#)

Henrik Grosskreutz

Session CP6 - Time Series and Sequence Analysis

295 [Simplex Distributions for Embedding Data Matrices over Time](#)

Kristian Kersting, Mirwaes Wahabzada, Christoph Römer, Christian Thureau, Agim Ballvora, Uwe Rascher, Jens Leon, Christian Bauckhage and Lutz Pluemer

307 [Transformation Based Ensembles for Time Series Classification](#)

Anthony Bagnall, Luke Davis, Jon Hills and Jason Lines

319 [Mining Compressing Sequential Patterns](#)

Hoang Thanh Lam, Fabian Moerchen, Dmitriy Fradkin and Toon Calders

331 [Beam Methods for the Profile Hidden Markov Model](#)

Samuel J. Blasiak, Huzefa Rangwala and Kathryn Laskey

343 [Optimal Distance Estimation Between Compressed Data Series](#)

Nikolaos Freris, Michail Vlachos and Serdar Kozat

Session CP7 - Kernels and Classification

355 [Multi-Objective Multi-Label Classification](#)

Chuan Shi, Xiangnan Kong, Philip Yu and Bai Wang

367 [Bayesian Supervised Multilabel Learning with Coupled Embedding and Classification](#)

Mehmet Gönen

379 [Subtree Replacement in Decision Tree Simplification](#)

Salvatore Ruggieri

391 [A Distributed Kernel Summation Framework for General-Dimension Machine Learning](#)

Dongryeol Lee, Richard Vuduc and Alexander Gray

403 [Kernelized Probabilistic Matrix Factorization: Exploiting Graphs and Side Information](#)

Tinghui Zhou, Hanhuai Shan, Arindam Banerjee and Guillermo Sapiro

CP8 - Social Media and Graphs

415 [On Dynamic Link Inference in Heterogeneous Networks](#)

Charu Aggarwal, Yan Xie and Philip Yu

427 [A Framework for the Evaluation and Management of Network Centrality](#)

Vatche Ishakian, Dóra Erdos, Evimaria Terzi and Azer Bestavros

439 [PICS: Parameter-free Identification of Cohesive Subgroups in Large Attributed Graphs](#)

Leman Akoglu, Hanghang Tong, Brendan Meeder and Christos Faloutsos

451 [Structural Analysis in Multi-Relational Social Networks](#)

Bing Tian Dai, Freddy Chong Tat Chua and Ee-Peng Lim

463[Influence Blocking Maximization in Social Networks under the Competitive Linear Threshold Model](#)

Xinran He, Guojie Song, Wei Chen and Qingye Jiang

Session - CP9 - Feature Selection, Networks and Prediction

475[Feature Selection over Distributed Data Streams through Convex Optimization](#)

Jacob Kogan

485[Feature Selection "Tomography" – Illustrating that Optimal Feature Filtering is Hopelessly Ungeneralizable](#)

George Forman

494[Sampling Strategies to Evaluate the Performance of Unknown Predictors](#)

Hamed Valizadegan, Saeed Amizadeh and Milos Hauskrecht

506[A Bayesian Markov-switching Model for Sparse Dynamic Network Estimation](#)

Huijing Jiang, Aurelie Lozano and Fei Liu

516[Learning Hierarchical Relationships among Partially Ordered Objects with Heterogeneous Attributes and Links](#)

Chi Wang, Jiawei Han, Qi Li, Xiang Li, Wen-Pin Lin and Heng Ji

Session CP10 - Transfer Learning

528[Transfer Learning of Distance Metrics by Cross-Domain Metric Sampling across Heterogeneous Spaces](#)

Guo-Jun Qi, Charu Aggarwal and Thomas Huang

540[Dual Transfer Learning](#)

Mingsheng Long, Jianmin Wang, Guiguang Ding, Wei Cheng, Xiang Zhang and Wei Wang

552[Transfer Significant Subgraphs across Graph Databases](#)

Xiaoxiao Shi, Xiangnan Kong and Philip Yu

564 [Transfer Topic Modeling with Ease and Scalability](#)

Jeon-Hyung Kang, Jun Ma and Yan Liu

Session CP11 - Applications - Healthcare and Networks

576 [SOR: Scalable Orthogonal Regression for Low-Redundancy Feature Selection and its Healthcare Applications](#)

Dijun Luo, Fei Wang, Jimeng Sun, Marianthi Markatou, Jianying Hu and Shahram Ebadollahi

588 [Mining Massive Archives of Mice Sounds with Symbolized Representations](#)

Jesin Zakaria, Sarah Rotschafer, Abdullah Mueen, Khaleel Razak and Eamonn Keogh

600 [IntruMine: Mining Intruders in Untrustworthy Data of Cyber-physical Systems](#)

Lu-An Tang, Quanquan Gu, Xiao Yu, Jiawei Han, Thomas La Porta, Alice Leung, Tarek Abdelzaher and Lance Kaplan

612 [Robust Reputation-Based Ranking on Bipartite Rating Networks](#)

Rong-Hua Li, Jeffrey Xu Yu, Xin Huang and Hong Cheng

Short Presentations

624 [Event Detection in Social Streams](#)

Charu Aggarwal and Karthik Subbian

636 [On Influential Node Discovery in Dynamic Social Networks](#)

Charu Aggarwal, Shuyang Lin and Philip Yu

648 [Query-based Biclustering using Formal Concept Analysis](#)

Faris Alqadah, Joel S. Bader, Rajul Anand and Chandan K. Reddy

660 [Granger Causality Analysis in Irregular Time Series](#)

Mohammad Taha Bahadori and Yan Liu

672 [Clustering Based on Yukawa Potential](#)

Xue Bai, Zezhen Lin, Yun Xiong and Yangyong Zhu

684[Deterministic CUR for Improved Large-Scale Data Analysis: An Empirical Study](#)

Christian Thureau, Kristian Kersting and Christian Bauckhage

696[Combining Active Learning and Dynamic Dimensionality Reduction](#)

Mustafa Bilgic

708[Context-aware Search for Personal Information Management Systems](#)

Jidong Chen, Wentao Wu, Hang Guo and Wei Wang

720[Mining Social Dependencies in Dynamic Interaction Networks](#)

Freddy Chong Tat Chua, Hady Lauw and Ee-Peng Lim

732[Detecting Irregularly Shaped Significant Spatial and Spatio-Temporal Clusters](#)

Weishan Dong, Xin Zhang, Li Li, Changhua Sun, Lei Shi and Wei Sun

744[Contextual Collaborative Filtering via Hierarchical Matrix Factorization](#)

Erheng Zhong, Wei Fan and Qiang Yang

756[Active Learning with Monotonicity Constraints](#)

Nicola Barile and Ad Feelders

768[Pseudo Cold Start Link Prediction with Multiple Sources in Social Networks](#)

Liang Ge and Aidong Zhang

780[Discovering Context-aware Influential Objects](#)

Yangpai Liu, Huiping Cao, Yifan Hao, Peng Han and Xinda Zeng

792[Monitoring and Mining Insect Sounds in Visual Space](#)

Yuan Hao, Bilson Campana and Eamonn Keogh

804[Image Mining of Historical Manuscripts to Establish Provenance](#)

Bing Hu, Thanawin Rakthanmanon, Bilson Campana, Abdullah Mueen and Eamonn Keogh

816[RP-growth: Top-k Mining of Relevant Patterns with Minimum Support Raising](#)

Yoshitaka Kameya and Taisuke Sato

828[Fast Random Walk Graph Kernel](#)

U Kang, Hanghang Tong and Jimeng Sun

839[Tracking Spatio-Temporal Diffusion in Climate Data](#)

Jaya Kawale, Aditya Pal and Rob Fatland

851[Group Sparsity in Nonnegative Matrix Factorization](#)

Jingu Kim, Renato Monteiro and Haesun Park

863[Global Linear Neighborhoods for Efficient Label Propagation](#)

Ze Tian and Rui Kuang

873[Generalized Similarity Kernels for Efficient Sequence Classification](#)

Pavel Kuksa, Imdadullah Khan and Vladimir Pavlovic

883[Detecting Extreme Rank Anomalous Collections](#)

Hanbo Dai, Feida Zhu, Ee-Peng Lim and Hwee Hwa Pang

895[Visualizing Variable-Length Time Series Motifs](#)

Yuan Li, Jessica Lin and Tim Oates

907[Which Distance Metric is Right: An Evolutionary K-Means View](#)

Chuanren Liu, Tianming Hu, Yong Ge and Hui Xiong

919[Constructing Training Sets for Outlier Detection](#)

Li-Ping Liu and Xiaoli Fern

930[A Flexible Open-Source Toolbox for Scalable Complex Graph Analysis](#)

Adam Lugowski, David Alber, Aydin Buluc, John Gilbert, Steve Reinhardt, Yun Teng and Andrew Waranis

942[Fast Robustness Estimation in Large Social Graphs: Communities and Anomaly Detection](#)

Fragkiskos D. Malliaros, Vasileios Megalooikonomou and Christos Faloutsos

954 [On Finding Joint Subspace Boolean Matrix Factorizations](#)

Pauli Miettinen

966 [Generalized Optimization Framework for Graph-based Semi-supervised Learning](#)

Marina Sokol, Konstantin Avratchenkov, Paulo Goncalves and Alexey Mishenin

975 [A Tree-Based Kernel for Graphs](#)

Giovanni Da San Martino, Nicolò Navarin and Alessandro Sperduti

987 [Density-based Projected Clustering over High Dimensional Data Streams](#)

Irene Ntoutsi, Arthur Zimek, Themis Palpanas, Peer Kröger and Hans-Peter Kriegel

999 [A Novel Approximation to Dynamic Time Warping allows Anytime Clustering of Massive Time Series Datasets](#)

Qiang Zhu, Gustavo Batista, Thanawin Rakthanmanon and Eamonn Keogh

1011 [Nearest-Neighbor Search on a Time Budget via Max-Margin Trees](#)

Parikshit Ram, Dongryeol Lee and Alexander Gray

1023 [Efficient Clustering of Metagenomic Sequences using Locality Sensitive Hashing](#)

Zeehasham Rasheed, Huzefa Rangwala and Daniel Barbara

1035 [Balancing Prediction and Recommendation Accuracy: Hierarchical Latent Factors for Preference Data](#)

Nicola Barbieri, Giuseppe Manco, Riccardo Ortale and Ettore Ritacco

1047 [On Evaluation of Outlier Rankings and Outlier Scores](#)

Erich Schubert, Remigius Wojdanowski, Arthur Zimek and Hans-Peter Kriegel

1059 [Regularized Structured Output Learning with Partial Labels](#)

Sundararajan Sellamanickam, Charu Tiwari and Sathiya Keerthi Selvaraj

1071 [The Similarity Between Stochastic Kronecker and Chung-Lu Graph Models](#)

C. Seshadhri, Ali Pinar and Tamara G. Kolda

1083 [WIGM: Discovery of Subgraph Patterns in a Large Weighted Graph](#)

Jiong Yang, Wei Su, Shirong Li and Mehmet Dalkilic

1095 [Legislative Prediction via Random Walks over a Heterogeneous Graph](#)

Jun Wang, Kush Varshney and Aleksandra Mojsilovic

1107 [An Iterative and Re-weighting Framework for Rejection and Uncertainty Resolution in Crowdsourcing](#)

Sihong Xie, Wei Fan and Philip S Yu

1119 [Citation Prediction in Heterogeneous Bibliographic Networks](#)

Xiao Yu, Quanquan Gu, Mianwei Zhou and Jiawei Han

1131 [Mining Multi-Label Data Streams Using Ensemble-Based Active Learning](#)

Peng Wang, Peng Zhang and Li Guo

1141 [Feature Selection for High-dimensional Integrated Data](#)

Charles Zheng, Scott Schwartz, Robert Chapkin, Raymond Carroll and Ivan Ivanov