

## Message from the PC Chairs

Welcome to the 2015 SIAM International Conference on Data Mining (SDM 2015)! In its fifteenth year, this conference continues to serve as an international forum for data mining researchers, students, practitioners and developers to exchange cutting-edge ideas, techniques, and experience.

This year we received a record 491 submissions. We recruited 154 Program Committee (PC) members along with 35 Senior Program Committee (SPC) members to help us with the review process. Each Program Committee member was assigned to at most ten papers while each Senior Program Committee member oversaw the review progress of twelve to fifteen papers.

A paper bidding phase was initiated among the reviewers once the submission deadline passed. Each submitted paper was then assigned based on their bids and subject areas to three Program Committee members to review. After the review deadline, the SPCs initiated and led discussions among the PC members. They then provided a final decision and a summary reflecting the reviews and discussions on their assigned papers. We evaluated these reviews and discussed with the SPCs regarding conflicting reviews. At the end of this process, 72 papers (14.66%) that we felt were of the highest quality were selected as full papers and another 36 papers (7.33%) were selected as posters. The overall acceptance rate is 22%. Many of the poster papers deserved full length presentation, but we were unable to accommodate them in the program. The oral presentation sessions, poster session, doctoral forum and tutorials give us a very full two and a half day technical program in addition to the final half day devoted to workshops.

We thank all the Program Committee members and external reviewers for their expert help with the challenging task of reviewing, discussing, and recommending papers. We particularly acknowledge the excellent help from the Senior Program Committee members: Aristides Gionis, Balaraman Ravindran, Carlotta Domeniconi, Diane Cook, Eammon Keogh, Evimaria Terzi, Wei Fan, Fosca Gianotti, Francesco Bonchi, George Karypis, Haesun Park, Hanghang Tong, Huan Liu, Hui Xiong, James Bailey, Jennifer Neville, Jian Pei, Jiawei Han, Jingrui He, Jun Liu, Jure Leskovec, Laks Lakshmanan, Minos Garofalakis, Myra Spiliopoulou, Philip Yu, Rajmonda Caceras, Sergei Vassilvitskii, Shivani Agarwal, Shuiwang Ji, Tamás Sarlós, Tanya Berger-Wolf, Tina Eliassi-Rad, Fei Wang, Wei Wang, Zhoujun Li. The conference could not occur without their help and the help of the many reviewers who so diligently undertook their duties.

We are sincerely grateful to all the SIAM staff members whose efforts have made our lives much easier. We thank Srinivasan Parthasarthy, Ke Wang, and Mohammed Zaki for sharing their experience and insights, for guiding us and providing excellent support and timely suggestions on many important issues. Perhaps most importantly, we thank the authors and attendees for this exchange of ideas and knowledge which is what SDM is all about.

We (the program committee, senior program committee, external reviewers, and ourselves) all dedicated much time to ensure the best possible program was put forward and we hope you enjoy the conference!

Suresh Venkatasubramanian and Jieping Ye  
Program Co-Chairs

## Table of Contents

### Session 1: Networks, Graphs I

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

### Session 2: Metric Learning, Feature Selection/Extraction

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

### Session 3: Clustering

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

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

## **Session 4: Applications**

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

## **Session 5: Recommendation, Classification**

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

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

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

## **Session 6: Security/Privacy, Social Media**

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

<b>Attacking DBSCAN for Fun and Profit.....</b>	<b>235</b>
Jonathan Crussell, Philip Kegelmeyer	

<b>Result Integrity Verification of Outsourced Privacy-preserving Frequent Itemset Mining .....</b>	<b>244</b>
Ruilin Liu, Wendy Wang	

<b>Modeling Users' Adoption Behaviors with Social Selection and Influence .....</b>	<b>253</b>
Ziqi Liu, Fei Wang, Qinghua Zheng	

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

## **Session 7: Time Series, Online Learning**

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

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

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

<b>Fast Mining of a Network of Coevolving Time Series .....</b>	<b>298</b>
Yongjie Cai, Hanghang Tong, Wei Fan, Ping Ji	

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

## Session 8: Matrix/Tensor

<b>Low Rank Representation on Riemannian Manifold of Symmetric Positive Definite Matrices .....</b>	<b>316</b>
Yifan Fu, Junbin Gao, Xia Hong, David Tien	
<b>Getting to Know the Unknown Unknowns: Destructive-Noise Resistant Boolean Matrix Factorization .....</b>	<b>325</b>
Sanjar Karaev, Pauli Miettinen, Jilles Vreeken	
<b>Convex Matrix Completion: A Trace-Ball Optimization Perspective .....</b>	<b>334</b>
Guangxiang Zeng, Ping Luo, Enhong Chen, Hui Xiong, Hengshu Zhu, Qi Liu	
<b>Near-separable Non-negative Matrix Factorization with <math>\ell_1</math>- and Bregman Loss Functions .....</b>	<b>343</b>
Abhishek Kumar, Vikas Sindhwani	
<b>Personalized TV Recommendation with Mixture Probabilistic Matrix Factorization .....</b>	<b>352</b>
Huayu Li, Hengshu Zhu, Yong Ge, Yanjie Fu, Yuan Ge	

## Session 9: Multi-source and Heterogeneous Learning

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

## Session 10: Networks, Graphs II

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

<b>Hidden Hazards: Finding Missing Nodes in Large Graph Epidemics.....</b>	<b>415</b>
Shashidhar Sundareisan, Jilles Vreeken, B. Aditya Prakash	

<b>Clustering and Ranking in Heterogeneous Information Networks via Gamma-Poisson Model.....</b>	<b>424</b>
Junxiang Chen, Wei Dai, Yizhou Sun, Jennifer Dy	

<b>A Divide-and-Conquer Algorithm for Betweenness Centrality.....</b>	<b>433</b>
Dora Erdos, Vatche Ishakian, Azer Bestavros, Evimaria Terzi	

<b>Frameworks to Encode User Preferences for Inferring Topic-sensitive Information Networks .....</b>	<b>442</b>
Qingbo Hu, Sihong Xie, Shuyang Lin, Wei Fan, Philip Yu	

## Session 11: Optimization

<b>Dropout Training of Matrix Factorization and Autoencoders for Link Prediction in Sparse Graphs.....</b>	<b>451</b>
Shuangfei Zhai, Zhongfei (Mark) Zhang	

<b>An ADMM Algorithm for Clustering Partially Observed Networks.....</b>	<b>460</b>
Necdet Serhat Aybat, Sahar Zarmehri, Soundar Kumara	

<b>Scaling log-linear analysis to datasets with thousands of variables .....</b>	<b>469</b>
Francois Petitjean, Geoffrey Webb	

<b>A Distributed Frank-Wolfe Algorithm for Communication-Efficient Sparse Learning.....</b>	<b>478</b>
Aurélien Bellet, Yingyu Liang, Alireza Bagheri Garakani, Maria-Florina Balcan, Fei Sha	

<b>Exceptional Model Mining with Tree-Constrained Gradient Ascent .....</b>	<b>487</b>
Thomas E. Krak, Ad Feelders	

## Session 12: Multi-task/Transfer Learning

<b>Formula: FactORized MUlti-task LeArning for task discovery in personalized medical models .....</b>	<b>496</b>
Jianpeng Xu, Jiayu Zhou, Pang-Ning Tan	

<b>Active Multi-task Learning via Bandits.....</b>	<b>505</b>
Meng Fang, Dacheng Tao	

<b>Hierarchical Active Transfer Learning .....</b>	<b>514</b>
David Kale, Marjan Ghazvininejad, Anil Ramakrishna, Jingrui He, Yan Liu	

<b>Learning Complex Rare Categories with Dual Heterogeneity .....</b>	<b>523</b>
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



<b>Principled Neuro-Functional Connectivity Discovery .....</b>	<b>631</b>
Kejun Huang, Nicholas Sidiropoulos, Evangelos Papalexakis, Christos Faloutsos, Partha Talukdar, Tom Mitchell	

<b>Estimating Ad Impact on Clicker Conversions for Causal Attribution: A Potential Outcomes Approach .....</b>	<b>640</b>
Joel Barajas, Ram Akella, Aaron Flores, Marius Holtan	

## Posters

<b>Towards Classification of Social Streams .....</b>	<b>649</b>
Min-Hsuan Tsai, Charu Aggarwal, Thomas Huang	

<b>Mobile App Security Risk Assessment: A Crowdsourcing Ranking Approach from User Comments.....</b>	<b>658</b>
Lei Cen, Deguang Kong, Hongxia Jin, Luo Si	

<b>Learning Compressive Sensing Models for Big Spatio-Temporal Data .....</b>	<b>667</b>
Dongun Lee, Jaesik Choi	

<b>Learning Stroke Treatment Progression Models for an MDP Clinical Decision Support System .....</b>	<b>676</b>
Dan C. Coroian, Kris Hauser	

<b>OnlineCM: Online Consensus Maximization with Missing Values.....</b>	<b>685</b>
Bowen Dong, Sihong Xie, Jing Gao, Wei Fan, Philip S. Yu	

<b>MET: A Fast Algorithm for Minimizing Propagation in Large Graphs with Small Eigen-Gaps .....</b>	<b>694</b>
Long Le, Tina Eliassi-Rad, Hanghang Tong	

<b>What shall I share and with Whom? - A Multi-Task Learning Formulation using Multi-Faceted Task Relationships.....</b>	<b>703</b>
Sunil Gupta, Santu Rana, Dinh Phung, Svetha Venkatesh	

<b>A Generalized Mixture Framework for Multi-label Classification.....</b>	<b>712</b>
Charmgil Hong, Iyad Batal, Milos Hauskrecht	

<b>Domain-Knowledge Driven Cognitive Degradation Modeling for Alzheimer's Disease .....</b>	<b>721</b>
Ying Lin, Kaibo Liu, Eunshin Byon, Xiaoning Qian, Shuai Huang	

<b>Ensemble Learning Methods for Binary Classification with Multi-modality within the Classes .....</b>	<b>730</b>
Anuj Karpatne, Ankush Khandelwal, Vipin Kumar	

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

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