AMiner: Search and Mining of Academic Social Networks (Postprint)

Authors: Wan, Huaiyu, Zhang, Yutao, Zhang, Jing, Tang, Jie, Tang, Jie

Date: 2022-11-25T00:00:00+00:00

Abstract

AMiner is a novel online academic search and mining system, and it aims to provide a systematic modeling approach to help researchers and scientists gain a deeper understanding of the large and heterogeneous networks formed by authors, papers, conferences, journals and organizations. The system is subsequently able to extract researchers' profiles automatically from the Web and integrates them with published papers by a way of a process that first performs name disambiguation. Then a generative probabilistic model is devised to simultaneously model the different entities while providing a topic-level expertise search. In addition, AMiner offers a set of researcher-centered functions, including social influence analysis, relationship mining, collaboration recommendation, similarity analysis and community evolution. The system has been in operation since 2006 and has been accessed from more than 8 million independent IP addresses residing in more than 200 countries and regions.

Full Text

Preamble

AMiner: Search and Mining of Academic Social Networks

Huaiyu Wan[1], Yutao Zhang[2], Jing Zhang[3] & Jie Tang[2]†

[1]Department of Computer Science, Beijing Jiaotong University, Beijing 100044, China
[2]Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[3]School of Information, Renmin University of China, Beijing 100872, China

Keywords: Academic social networks; Profile extraction; Name disambiguation; Topic modeling; Expertise search; Network mining

Citation: H. Wan, Y. Zhang, J. Zhang & J. Tang. AMiner: Search and mining of academic social networks. *Data Intelligence* 1(2019), 58-76. doi: 10.1162/dint_a_{00006}
Received: May 22, 2018; **Revised:** June 8, 2018; **Accepted:** June 12, 2018

Abstract

AMiner is a novel online academic search and mining system designed to provide a systematic modeling approach that helps researchers and scientists gain deeper understanding of large, heterogeneous networks formed by authors, papers, conferences, journals, and organizations. The system automatically extracts researcher profiles from the Web and integrates them with published papers through a name disambiguation process. It then employs a generative probabilistic model to simultaneously model different entities while enabling topic-level expertise search. Additionally, AMiner offers researcher-centered functions including social influence analysis, relationship mining, collaboration recommendation, similarity analysis, and community evolution. Operational since 2006, the system has been accessed from over 8 million unique IP addresses across more than 200 countries and regions.

† Corresponding author: Jie Tang (Email: jietang@tsinghua.edu.cn; ORCID: 0000-0002-6882-4044)

1. Introduction

Academic social networking platforms such as Google Scholar, Microsoft Academic, Semantic Scholar, ResearchGate, and Academia.edu have gained tremendous popularity over the past decade. These systems share a common goal: providing researchers with integrated platforms to query academic information, share achievements, and connect with peers.

While these systems have investigated various issues within academic social networks, most approaches treat problems separately through independent processes. Consequently, there is no coherent framework for mining entire academic social networks. This gap stems from two primary challenges. First, there is a lack of semantic-based information: user profiles obtained solely through manual entry or heuristic extraction are often incomplete or inconsistent, as users may be unwilling to provide personal details. Second, there is no unified modeling approach for effective social network mining. Traditional methods model different information sources in academic networks individually, failing to capture dependencies between them. However, such dependencies exist within social data, and high-quality search services must account for the intrinsic relationships between heterogeneous information sources.

AMiner, the second generation of the ArnetMiner system, is designed to search and mine academic publications on the Internet using social network analysis to identify connections between researchers, conferences, and publications. AMiner addresses four fundamental questions: (1) How can researcher profiles be automatically extracted from the existing Web? (2) How can extracted information (researcher profiles and publications) from different sources be integrated? (3) How can different types of information sources be modeled within a unified framework? (4) How can enhanced search services be provided within the constructed network?

To answer these questions, AMiner implements a series of novel approaches. The overall architecture, shown in Figure 1 [Figure 1: see original paper], comprises five main components. First, the **Extraction** component automatically extracts researcher profiles from the Web by collecting and identifying relevant pages (e.g., homepages or biography pages), then applies a unified approach to extract data from these documents. It also extracts publications from online digital libraries using heuristic rules, and employs a simple yet effective big data approach for profiling Web users. Second, the **Integration** component joins and integrates extracted researcher profiles with publications, using researcher names as identifiers. To address name ambiguity, a probabilistic model and comprehensive framework have been developed. Integrated data are then stored, sorted, and indexed into a researcher network knowledge base.

Third, **Storage and Access** provides storage and indexing for the extracted and integrated data. The system uses Jena for storing and retrieving ontological data and employs inverted file indexing to facilitate information retrieval. Fourth, the **Modeling** component utilizes a generative probabilistic model to simultaneously model different information sources, estimating mixture distributions of topics associated with each source. Finally, **Services** provides several powered functions based on the modeling results, including profile search, expert finding, conference analysis, course search, sub-graph search, topic browser, academic ranks, and user management.

For key features such as profile extraction, name disambiguation, academic topic modeling, expertise search, and academic social network mining, we propose new approaches to overcome limitations of conventional methods. The remainder of this paper is organized as follows: Section 2 discusses related work, Section 3 presents our proposed approaches, Section 4 demonstrates applications, Section 5 describes our datasets, and Section 6 concludes.

---

2. Related Work

Several academic social network issues have been investigated, and various systems have been developed. Google Scholar provides a search engine for identifying hyperlinks to publicly available publications or those accessible through institutional libraries. While not a social networking site in the traditional sense,

it has become an essential platform for searching academic resources, tracking research, promoting achievements, and monitoring impact. Registered users can create profiles to list research interests, manage publications, correct co-authors, and view annual citation metrics. Its social features are limited: users can follow researchers to receive email alerts about new publications or citations, and set up field-based alerts.

Microsoft Academic employs machine learning, semantic analysis, and data mining to help users explore academic information more powerfully. Users can create accounts and public profiles by claiming their publications. It offers extensive "follow" functions for researchers, publications, journals, conferences, organizations, and research topics. Based on publication history and followed events, Microsoft Academic displays relevant items on personalized homepages. Rather than simple keyword search, it provides relevant results and recommendations to help users discover more resources and support expansive research experiences.

Semantic Scholar is designed as a "smart" search engine to help researchers find better publications faster. It combines machine learning, natural language processing, and computer vision to analyze publications and extract important features, adding a semantic analysis layer to traditional citation analysis. Compared to Google Scholar and Microsoft Academic, Semantic Scholar quickly highlights important papers and identifies connections between them, providing influential citations, images, and key phrases that become highly relevant to users' work.

ResearchGate aims to connect geographically distant researchers and enable continuous communication. Users have profiles and can share research outputs including papers, data, chapters, patents, proposals, algorithms, presentations, and source code. They can follow others' activities and engage in discussions. Organized primarily around research topics, ResearchGate maintains its own ResearchGate Score based on content contributions, profile details, and site participation (e.g., asking and answering questions).

Academia.edu is a for-profit academic social networking site where users create profiles, share works, monitor impact, select interest areas, and follow research evolution in specific fields. Users can browse networks of similar-interest researchers worldwide. The platform includes an analytics dashboard showing real-time influence and diffusion of works, plus an alert service that emails users when followed researchers publish new papers or when followed topics have new activity, potentially raising paper awareness among potential citators.

Although these systems have integrated vast academic resources and provide abundant search and networking functions, they lack systematic semantic-level analysis or mining. In contrast, AMiner' s primary objective is to provide a unified modeling approach for deeper understanding of semantic connections in large, heterogeneous academic networks comprising authors, papers, conferences, journals, and organizations, thereby enabling topic-level expertise search

and researcher-centered functions.

---

3. Methodology

This section details the challenges AMiner addresses and presents our methods and solutions.

3.2 Name Disambiguation

We have collected over 200 million publications from online digital libraries including DBLP, ACM DL, CiteSeerX, and others. In each source, authors are identified by names, which we use as identifiers to integrate researcher profiles with publication data. This process inevitably creates ambiguity. Several years ago, we proposed a probabilistic framework based on Hidden Markov Random Fields (HMRF) that captures dependencies between observations (where each paper is an observation), casting disambiguation as assigning tags to papers where each tag represents an actual researcher. More recently, we developed an additional comprehensive framework that incorporates a novel representation learning method using both global and local information, plus an end-to-end cluster size estimation method. To improve accuracy, we involve human annotators in the disambiguation loop. This method now handles name disambiguation at billion-scale in AMiner, demonstrating its effectiveness and efficiency.

3.3 Topic Modeling

In academic search, representing text documents, author interests, and conference themes is critical. Traditional approaches use "bag of words" (BOW) representations, which cannot capture semantic dependencies between words. Additionally, academic search involves multiple information source types, making it challenging to model dependencies between them. Existing topic models such as probabilistic Latent Semantic Indexing (pLSI), Latent Dirichlet Allocation (LDA), and Author-Topic models cannot be directly applied because they fail to capture all intrinsic dependencies between papers and conferences.

We propose a unified topic modeling approach called the Author-Conference-Topic (ACT) model that simultaneously models characteristics of documents, authors, conferences, and their dependencies (using "conference" to denote conferences, journals, and books). Different strategies can model topic distributions, yielding different knowledge representation capacities. In the first variant, each author has a mixture of topic weights, with each word token and conference stamp generated from a sampled topic. In the second variant, each author-conference pair has topic mixture weights, with word tokens generated from sampled topics. In the third variant, each author is associated with topics, word tokens are generated from sampled topics, and conferences are generated from the topics of all word tokens in a paper.

3.4 Expertise Search

When searching for academic resources, users seek authors with specific expertise and relevant papers or conferences. AMiner presents a topic-level expertise search framework that differs from traditional document-level retrieval by studying expertise search at the topic level across heterogeneous networks. We propose a unified Citation-Tracing-Topic (CTT) model to simultaneously model topical aspects of different objects in the academic network. Based on learned topic models, we investigate expertise search from three dimensions: ranking, citation tracing analysis, and topic graph search. Specifically, we propose a topic-level random walk method for ranking objects, seek to uncover how studies influence follow-up work through citation tracing analysis, and have developed a topical graph search function based on topic modeling and citation tracing.

3.5 Academic Social Network Mining

AMiner provides researcher-centric mining functions including social influence analysis, relationship mining, similarity analysis, collaboration recommendation, and community evolution.

Social Influence Analysis. In large social networks, individuals are influenced by others for various reasons. We propose a Topic Affinity Propagation (TAP) model to differentiate and quantify social influence, which can use any topic modeling results and existing network structure for topic-level influence propagation. Recently, we designed an end-to-end framework called DeepInf for feature representation learning and social influence prediction. Each user is represented by their local sub-network embedding, and a graph neural network learns the sub-network representation, effectively integrating user-specific features and network structures.

Social Relationship Mining. Inferring relationship types between users is crucial. We propose a two-stage Time-constrained Probabilistic Factor Graph model (TPFG) to infer advisor-advisee relationships in co-author networks by decomposing the joint probability of unknown advisors across all authors. Additionally, we developed the TranFG framework for classifying social relationship types across heterogeneous networks, incorporating social theories into a factor graph model to improve prediction accuracy in target networks by borrowing knowledge from source networks.

Similarity Analysis. Estimating vertex similarity is fundamental in social network analysis. We propose a sampling-based method called Panther to estimate top-k similar vertices. Given a network, Panther randomly generates paths of pre-defined length, modeling similarity between two vertices as the probability they appear on the same paths.

Collaboration Recommendation. Interdisciplinary collaborations significantly impact society but are difficult to establish. Analyzing cross-domain collaboration data from publications, we propose a Cross-domain Topic Learn-

ing (CTL) model for collaboration recommendation. To handle sparse connections, CTL consolidates existing cross-domain collaborations through topic layers rather than author layers. To handle complementary expertise, CTL separately models topic distributions from source and target domains plus their correlations. To handle topic skewness, CTL models only topics relevant to cross-domain collaboration.

Community Evolution. Social networks are dynamic, making it interesting to study how people form clusters and how these clusters evolve over time. We study co-evolution of multi-typed objects in a special heterogeneous network type called star networks, examining how multi-typed objects influence each other during network evolution. We propose a Hierarchical Dirichlet Process Mixture Model-based evolution model that detects co-evolution of multi-typed objects as multi-typed cluster evolution in dynamic star networks, with an efficient inference algorithm for model learning.

---

4. Application

AMiner provides comprehensive search and mining services for researcher social networks, focusing on: (1) creating semantic-based researcher profiles by extracting information from the distributed Web; (2) integrating academic data (bibliographic data and researcher profiles) from multiple sources; (3) enabling accurate search in heterogeneous networks; and (4) analyzing and discovering interesting patterns in the constructed researcher social network.

Profile Search. Inputting a researcher name (e.g., Jie Tang) returns a semantic-based profile created using information extraction techniques. Profiles include contact information, photo, citation statistics, academic achievement evaluation, (temporal) research interests, educational history, personal social graph, research funding (currently US and China only), and publication records with citation information automatically assigned to different domains.

Expert Finding. Inputting a query (e.g., "data mining") returns experts on that topic, along with top conferences and ranked papers. Two ranking algorithms are available: VSM (similar to conventional language models) and ACT (based on our Author-Conference-Topic model). Users can provide feedback on search results.

Conference Analysis. Inputting a conference name (e.g., KDD) returns the most active researchers in that conference and top-ranked papers.

Course Search. Inputting a query (e.g., "data mining") returns researchers teaching relevant courses.

Sub-Graph Search. Inputting a query (e.g., "data mining") first identifies relevant topics (e.g., "Data Mining," "XML Data," "Data Mining/Query Processing," "Web Data/Database Design," "Web Mining"), then displays the most

important sub-graph discovered for each topic with a summary.

Topic Browser. Based on our ACT model, we automatically discover 200 hot topics from publications, assigning labels to represent their meanings. The browser presents the most active researchers, relevant conferences/papers, and topic evolution trends.

Academic Ranks. We define eight measures to evaluate researcher achievement: h-index, citations, uptrend, activity, longevity, diversity, sociability, and new star. For each measure, we provide ranking lists across domains (e.g., highest-cited researchers in "data mining").

User Management. Registered users can: (1) modify extracted profile information; (2) provide search result feedback; (3) follow researchers; and (4) create AMiner pages for advertising conferences, workshops, or recruiting students.

---

5. Data Set

By June 2018, AMiner had collected a large scholar dataset with over 130 million researcher profiles and 233 million publications from the Internet, along with numerous subsets for different research purposes, available at https://www.aminer.cn/data.

Citation Network. Extracted from DBLP, ACM DL, and other sources, this dataset contains 1,572,277 papers and 2,084,019 citation relationships. Each paper includes abstract, authors, year, venue, and title. It supports clustering with network and side information, influence analysis, influential paper identification, and topic modeling.

Academic Social Network. This dataset includes papers, citations, author information, and collaborations, containing 1,712,433 authors, 2,092,356 papers, 8,024,869 citations, and 4,258,615 co-author relationships.

Advisor-Advisee. Comprising 815,946 authors and 2,792,833 co-author relationships, we created a ground truth subset for evaluating advisor-advisee inference by collecting data from the Mathematics Genealogy Project and AI Genealogy Project, plus manual crawling from researcher homepages, resulting in 1,534 labeled relationships (514 advisor-advisee).

Topic-Co-Author. A topic-based co-author network with 640,134 authors across eight topics and 1,554,643 co-author relationships. Topics include Data Mining/Association Rules, Web Services, Bayesian Networks/Belief Function, Web Mining/Information Fusion, Semantic Web/Description Logics, Machine Learning, Database Systems/XML Data, and Information Retrieval.

Topic-Paper-Author. Collected for cross-domain recommendation, this dataset contains 33,739 authors across five topics and 139,278 co-author relationships. Topics are Data Mining (6,282 authors, 22,862 relationships),

Medical Informatics (9,150 authors, 31,851 relationships), Theory (5,449 authors, 27,712 relationships), Visualization (5,268 authors, 19,261 relationships), and Database (7,590 authors, 37,592 relationships).

Topic-Citation. A topic-based citation network with 2,329,760 papers across 10 topics and 12,710,347 citation relationships. The 10 topics match those in Topic-Co-Author plus Pattern Recognition/Image Analysis and Natural Language System/Statistical Machine Translation.

Kernel Community. A co-authorship network with 822,415 nodes and 2,928,360 undirected edges, where vertices represent authors and edges represent co-author relationships.

Dynamic Co-Author. Contains 1,768,776 papers published from 1986 to 2012 with 1,629,217 authors. Each year is a timestamp (27 total). Edges connect authors who co-authored at least one paper in the most recent three years (including current year). Undirected edges are converted to two symmetric directed edges.

Expert Finding. A benchmark dataset containing 1,781 experts across 13 topics.

Association Search. Used to evaluate association search approaches, containing 8,369 author pairs across nine topics, each with source and target authors.

Topic Model Results for AMiner Dataset. ACT model results on the AMiner dataset containing top 1,000,000 papers and authors across 200 topics.

Co-Author. A co-author network from AMiner with 1,560,640 authors and 4,258,946 co-author relationships.

Disambiguation. A dataset for studying name disambiguation in digital libraries, containing 110 authors with affiliations and ground truth disambiguation results.

---

6. Conclusion

This paper presents AMiner, a novel online academic search and mining system and the second generation of ArnetMiner. We described the overall architecture with five main components: extraction, integration, storage and access, modeling, and services. We introduced key methodologies including profile extraction and user profiling methods, name disambiguation algorithms, topic modeling approaches, expertise search strategies, and academic social network mining methods. We also demonstrated typical applications and a broad offering of available datasets on the platform.

While AMiner remains in development regarding resource scale and service quality, we plan to exploit additional intelligent methods for mining deep knowledge

from scientific networks and deploy a more convenient, personalized framework for academic search and discovery services.

References

J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, & Z. Su. ArnetMiner: Extraction and mining of academic social networks. In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD' 08)*, 2008, pp. 990–998. doi: 10.1145/1401890.1402008.

J. Tang, D. Zhang, & L. Yao. Social network extraction of academic researchers. In: *Proceedings of 2007 IEEE International Conference on Data Mining (ICDM' 07)*, 2007, pp. 292–301. doi: 10.1109/ICDM.2007.30.

J. Tang, L. Yao, D. Zhang, & J. Zhang. A combination approach to Web user profiling. *ACM Transactions on Knowledge Discovery from Data* 5(1) 2010, Article No. 2. doi: 10.1145/1870096.1870098.

X. Gu, H. Yang, J. Tang, J. Zhang, F. Zhang, D. Liu, W. Hall, & X. Fu. Profiling Web users using big data. *Social Network Analysis and Mining* 8(1) 2018, Article No. 24. doi: 10.1007/s13278-018-0495-0.

J. Tang, A.C.M. Fong, B. Wang, & J. Zhang. A unified probabilistic framework for name disambiguation in digital library. *IEEE Transaction on Knowledge and Data Engineering* 24(6) 2012, 975–987. doi: 10.1109/TKDE.2011.13.

Y. Zhang, F. Zhang, P. Yao, & J. Tang. Name disambiguation in AMiner: Clustering, maintenance, and human in the loop. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD' 18)*, 2018, pp. 1002–1011. doi: 10.1145/3219819.3219859.

J.J. Carroll, I. Dickinson, C. Dollin, D. Reynolds, A. Seaborne, & K. Wilkinson. Jena: Implementing the semantic Web recommendations. In: *Proceedings of the 13th World Wide Web Conference (WWW' 04)*, 2004, pp. 74–83. doi: 10.1145/1013367.1013381.

C.J. van Rijsbergen. *Information retrieval*. London: Butterworths, 1979.

A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.J. Hsu, & K. Wang. An overview of Microsoft Academic Service (MA) and applications. In: *Proceedings of the 24th International Conference on World Wide Web (WWW' 15 Companion)*, 2015, pp. 243–246. doi: 10.1145/2740908.2742839.

D. Brickley, & L. Miller. FOAF vocabulary specification. Available at: http://xmlns.com/foaf/0.1/.

C. Cortes, & V. Vapnik. Support-vector networks. *Machine Learning* 20(3)(1995), 273–297. doi: 10.1007/BF00994018.

J.D. Lafferty, A. McCallum, & F.C.N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: *Proceedings of the 18th International Conference on Machine Learning (ICML' 01)*, 2001, pp. 282–289. Available at: http://portal.acm.org/citation.cfm?id=655813.

S. Basu, M. Bilenko, & R.J. Mooney. A probabilistic framework for semi-supervised clustering. In: *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD' 04)*, 2004, pp. 59–68. doi: 10.1145/1014052.1014062.

T. Hofmann. Probabilistic Latent Semantic indexing. In: *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR' 99)*, 1999, pp. 50–57. doi: 10.1145/312624.312649.

D.M. Blei, & J.D. McAuliffe. Supervised topic models. In: *Proceedings of the 19th Neural Information Processing Systems (NIPS' 07)*, 2007, pp. 121–128. Available at: http://papers.nips.cc/paper/3328-supervised-topic-models.

M. Rosen-Zvi, T. Griffiths, M. Steyvers, & P. Smyth. The author-topic model for authors and documents. In: *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence (UAI' 04)*, 2004, pp. 487–494. Available at: https://dl.acm.org/citation.cfm?id=1036902.

M. Steyvers, P. Smyth, & T. Griffiths. Probabilistic author-topic models for information discovery. In: *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD' 04)*, 2004, pp. 306–315. doi: 10.1145/1014052.1014087.

J. Tang, J. Zhang, R. Jin, Z. Yang, K. Cai, L. Zhang, & Z. Su. Topic level expertise search over heterogeneous networks. *Machine Learning Journal* 82(2)(2011), 211–237. doi: 10.1007/s10994-010-5212-9.

J. Tang, J. Sun, C. Wang, & Z. Yang. Social influence analysis in large-scale networks. In: *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD' 09)*, 2009, pp. 807–816. doi: 10.1145/1557019.1557108.

J. Qiu, J. Tang, H. Ma, Y. Dong, K. Wang, & J. Tang. DeepInf: Modeling influence locality in large social networks. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD' 18)*, 2018, pp. 2110–2119. Available at: https://www.haoma.io/pdf/deepinf.pdf.

C. Wang, J. Han, Y. Jia, J. Tang, D. Zhang, Y. Yu, & J. Guo. Mining advisor-advisee relationships from research publication networks. In: *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD' 10)*, 2010, pp. 203–212. doi: 10.1145/2910896.2925435.

J. Tang, T. Lou, J. Kleinberg, & S. Wu. Transfer learning to infer social ties across heterogeneous networks. *ACM Transactions on Information Systems* 34(2)(2016), Article No. 7. doi: 10.1145/2746230.

J. Zhang, J. Tang, C. Ma, H. Tong, Y. Jing, & J. Li. Panther: Fast top-k similarity search on large networks. In: *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD' 15)*, 2015, pp. 1445–1454. doi: 10.1145/2783258.2783267.

J. Tang, S. Wu, J. Sun, & H. Su. Cross-domain collaboration recommendation. In: *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD' 12)*, 2012, pp. 1285–1293. Available at: http://keg.cs.tsinghua.edu.cn/jietang/publications/KDD12-Tang-et-al-Cross-Domain-Collaboration-Recommendation.pdf.

Y. Sun, J. Tang, J. Han, C. Chen, & M. Gupta. Co-evolution of multi-typed objects in dynamic star networks. *IEEE Transaction on Knowledge and Data Engineering* 26(12)(2014), 2942–2955. doi: 10.1109/TKDE.2014.2334316.

Author Biography

Huaiyu Wan is an Associate Professor at the Department of Computer Science, Beijing Jiaotong University. He received his PhD from the School of Computer and Information Technology, Beijing Jiaotong University. His research interests include social network mining, user behavior analysis, and traffic data mining.

Yutao Zhang is a postdoctoral researcher at the Department of Computer Science and Technology, Tsinghua University. He received his PhD from the Department of Computer Science and Technology, Tsinghua University. His research interests include social network mining, text mining, and visual analytics.

Jing Zhang is an Assistant Professor at the Department of Computer Science and Technology, Information School, Renmin University of China. She received her PhD from the Department of Computer Science and Technology, Tsinghua University. Her research interests include social network mining, graph mining, text mining, and deep learning.

Jie Tang is an Associate Professor at the Department of Computer Science and Technology, Tsinghua University. His main research interests include data mining algorithms and social network theories. He has been a visiting scholar at Cornell University, Chinese University of Hong Kong, Hong Kong University of Science and Technology, and Leuven University. He has published over 100 research papers in major international journals such as *Machine Learning*, *ACM Transactions on Knowledge Discovery from Data (TKDD)*, and *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, and conferences including KDD, IJCAI, AAAI, ICML, WWW, SIGIR, SIGMOD, and ACL.

Note: Figure translations are in progress. See original paper for figures.