Research Statement: Data-Driven Behavioral Analytics with Networks

Meng Jiang, University of Illinois at Urbana-Champaign
http://www.meng-jiang.com

Behavior is defined as *interaction* made by individuals in conjunction with themselves or their environment[1]. Thanks to Information Technologies, the human behaviors are broadly recorded in an unprecedented level. This gives us an opportunity for getting insights of behaviors and our societies from large-scale real data. Dr. Robin Staffin, the Director for Basic Research in the Department of Defense (DoD), listed *Computational Modeling of Human Behavior* as one of the six high-priority topics in DoD Basic Research[2]. In order to provide a fundamental understanding and predictive capability of human behavior dynamics from individuals to societies, behavioral analysis has to face the following complexity of behvaior: (1) human behaviors are highly dependent on *social contexts*; (2) only by modeling *spatiotemporal contexts*, can we understand when, where and how the behavior happens; (3) *behavioral intentions* including monetary incentives and other suspicious purposes are a nontrivial part of behavior modeling; (4) user preference on the *content* is an important driving factor. Experience-driven approaches rely on expertise in developing features of intelligent and trustworthy systems (e.g., recommender systems, anti-fraud/anti-spam systems). Taking the advantage of massive, available behavioral data, I have been proposing and promoting data-driven approaches that leverage the methodology of *observation*, *representation* and *models* from real data beyond our eyes and hands. I focus on analyzing behavioral data with network models relying little to none of human annotations for themes including (T1) mining multidimensional behavior networks with social spatiotemporal contexts, (T2) structuring behavioral content into heterogeneous information networks of entities and attributes, and (T3) integrating behavior networks and information networks for in-depth behavioral analysis.

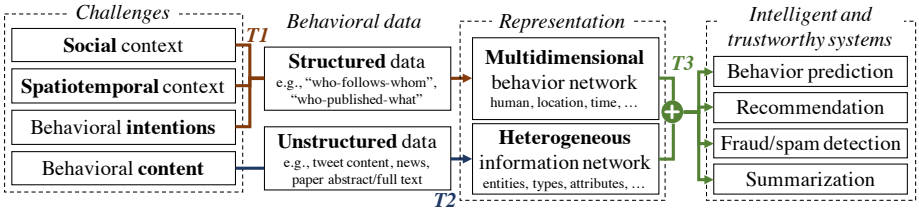


Figure 1: Towards building intelligent and trustworthy systems and addressing the four challenges, I have been proposing data-driven approaches of discovering rich knowledge from behavioral data by *mining* behavior networks (T1), *constructing* information networks (T2) and *integrating* them into in-depth behavioral analysis (T3).

I. THESIS AND POSTDOCTORAL WORK

My thesis titled "Modeling Complex Behaviors in Social Media" focuses on Theme T1 of understanding behaviors under social spatiotemporal contexts, and my two-year postdoctoral research is broken down to T2 and T3 with contributions mostly in in-depth behavioral analysis with unstructured data. The three tasks form the full picture of "Data-Driven Behavioral Analytics with Networks".

T1. Mining behavior network with social spatiotemporal contexts

During my five-year Ph.D., I feel fortunate joining the long-term collaboration between Tencent Weibo (one of the largest Twitter-style social platforms in China) and Tsinghua University, which keeps my hands on massive real data generated by millions of users and billions of social relations and real-world problems including building social recommender systems and anti-fraud systems.

T1.1 Modeling social spatiotemporal contexts for behavior prediction

Weibo suffers from *low* conversion rate: users received too much (interestingless) information from crowds and generated fewer than 6 retweets within every 100 news feed requests. Can we recommend tweets by predicting their behaviors to address the issue of information overload? Although collaborative filtering techniques have been widely used, we know little about why users adopt or reject items, which has been the bottleneck for further improving the recommending performance. How to fully exploit social contextual information from "who-posts-what" behavior network and "who-follows-whom" social network to reveal the information adoption mechanism, and accordingly propose a method for recommendation, is rather challenging and demanded.

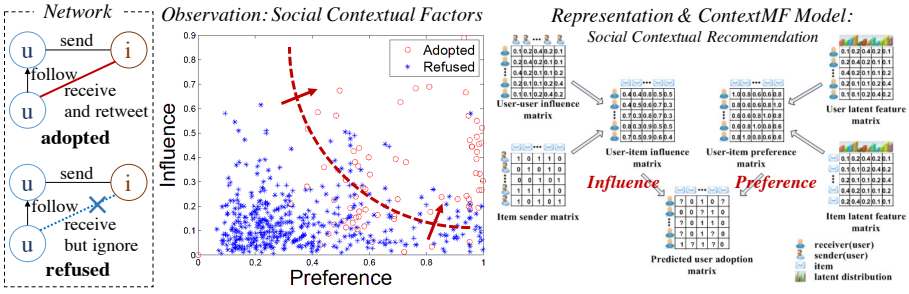


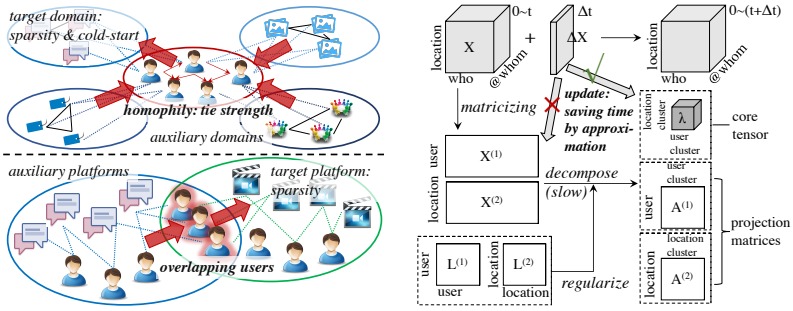
Figure 2: Uncovering two significant factors of user decisions with data, we proposed CONTEXTMF [1] that integrates preference and influence in social recommendation.

To achieve this goal, we observe from the real network data that personal preference on the content and interpersonal influence from the sender are two significant factors that determine users' decisions (see Figure 2). We propose a probabilistic matrix factorization model, CONTEXTMF [1], to fuse the behaviors and social contexts. In this model, there are three low-rank latent spaces respectively corresponding to the user space, item space and influence space. They are regularized by the observed preference similarity matrix, content similarity matrix and interaction frequency matrix. The product of the user space and item space corresponds to the

[1] Behavior – Wikipedia, the free encyclopedia: https://en.wikipedia.org/wiki/Behavior
[2] National Defense Industrial Association (NDIA) Conference Proceedings: http://www.dtic.mil/ndia/

preference; the product of influence and preference is proportional to the probability of item adoption. Empirical results on two real social datasets (one from Weibo) demonstrate that CONTEXTMF increases the accuracy of prediction by 21% and 17%, respectively.

Cross-domain behavior modeling. The principle is that if we have denser user-item links of behavioral data, we acquire more knowledge and thus predict more accurately. However, the real case comes to be sparsity and cold start (when the user has not yet adopted or rejected any item). Fortunately, the social platform enables users generating content in multiple domains such as profiling tags, images and interest groups. We propose Hybrid Random Walk (HYBRIDRW) [2] to transfer knowledge across multiple domains on a star-structured network (see Figure 3) for social recommendation. HYBRIDRW collaboratively integrates multiple domains to discover the common knowledge of user tie strength in the central social domain and accordingly alleviates the sparsity issue in individual domains. Empirical results show that given profiling tag information, our method needs only 35% training data of the target domain to achieve the same performance of taking all the training data but no profiling tag.



(a) HYBRIDRW and XPTRANS (b) FEMA

Figure 3: Contextual behavior modeling: (a) bridging behaviors across domains and platforms to alleviate the issue of sparsity, (b) modeling spatiotemporal information with Flexible Evolutionary Multifaceted Analysis.

Cross-platform behavior modeling. A more important issue is that emerging platforms meet more serious sparsity and cold start problems. Can we utilize rich social platform data to improve the accuracy of predicting behaviors in other platforms (e.g., movie ratings in Netflix, ride requests in Uber)? The natural bridge across platforms is the small number of partially overlapping users, for example, users can register Uber with their Facebook accounts. How to make full use of the bridge is an open and challenging problem. We propose XPTRANS [3] that transfers knowledge across platforms by jointly optimizing different user and item spaces on different platforms and using pairwise similarity of the overlapping users to constrain the spaces. By transferring via the little percentage (i.e., 1.1%) of the overlapping users, the predicting performance on the non-overlapping users is even better than that on the overlapping users (who generate more data) without transfer.

Spatiotemporal behavior modeling. Human behavior is a product of and evolves with the changing of a multitude of interrelated factors such as physical environment, social identity and interaction. We represent behaviors of multiple dimensions of user who posts, user who is mentioned, word and location as high-order tensor sequences. We propose a Flexible Evolutionary Multi-faceted Analysis (FEMA) method [4] based on a dynamic scheme of tensor factorization in which flexible regularizers of social-relationship and location-distance information are imposed to alleviate the high sparsity. To address the complexity issue, we give approximation algorithms based on Tensor Perturbation theory to factorize the updated tensor with sparse increments, where the loss bound is theoretically proved. Experimental results demonstrate that our method achieves 30.8% higher accuracy when it considers multi-dimensions and 17.4% higher accuracy when it uses the regularizers. Moreover, it can reduce the time cost from hours to minutes.

Impact.

- CONTEXTMF [1] and HYBRIDRW [2] are the 3^{rd} and 9^{th} most cited papers of CIKM 2012 with **119** and **47** citations among 146 accepted papers at the acceptance rate of 13.4%, where as the median number of citations for CIKM 2012 is 11. Adding the feature of scalability, CONTEXTMF+ [5] has appeared in TKDE 2014 with **35** citations.
- CONTEXTMF+ has been deployed in Weibo's recommender system. Online testing demonstrates that the conversion rate is improved from 5.78% to 8.27% (relatively **43%**). Note that the rate is strongly related to the company's ad revenue.

T1.2 Modeling social spatiotemporal contexts for suspicious behavior detection

Given "who-follows-whom" networks, how can we automatically detect fake followers with high recall? The fraudsters are paid to make certain accounts seem more legitimate or famous through giving them many additional followers. Twitter admitted 5% of its users are fake[3]; Weibo suffered the same and more serious problem. Experience-driven approaches learn classifiers with features such as the numbers of followees, hashtags and URLs. However, the fraudsters are smart: they don't have to follow or publish many; instead, a large set of them consistently connect to the customers.

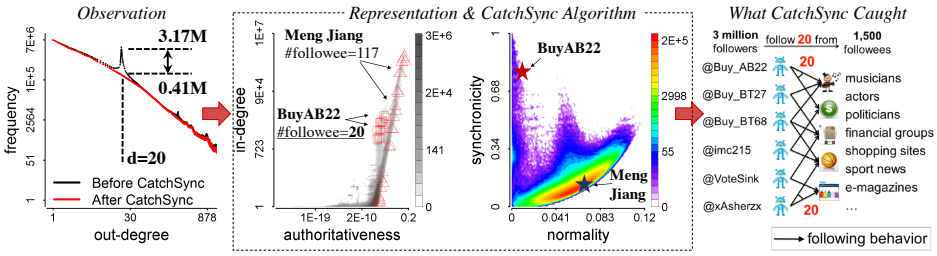


Figure 4: CATCHSYNC [6] spots that fake followers consistently connect to customers of similar features: their synchronized behaviors create spikes on degree distributions.

Essentially, we reconsider the detection problem by revealing the fraudsters' manipulation on the network. As shown in Figure 4, the spikes on the out-degree distribution indicate millions of anomalous nodes on the network. The fraudsters exhibit behavior that is

[3]Business Insider. http://www.businessinsider.com/5-of-twitter-monthly-active-users-are-fake-2013-10

(1) synchronized (i.e., cause to occur at the same rate): they often connect to the very same 20, 100 or 500 targets, and (2) abnormal: their behavior pattern is different from the majority of nodes. We propose a scalable and parameter-free algorithm, CATCHSYNC [6], to measure the two properties (synchronicity and normality) of groups of followees. We prove that the synchronicity has a parabolic lower bound of the normality. CATCHSYNC detects millions of fake followers who have unexpectedly high synchronicity and successfully recovers the distribution into a power-law shape, which demonstrates high recall of the performance.

Evaluating suspiciousness across dimensions. Besides inflating the number of fans, Weibo's fraudsters are trying to manipulate the popularity of trending topics. So which seems more suspicious: 5,000 tweets from 200 users on 5 IP addresses, or 10,000 tweets from 500 users on 500 IP addresses but all with the same hashtag and all in 10 minutes? The literature has many methods that find dense blocks in matrices and tensors, but no method gives a principled way to score the suspiciousness of dense blocks with different numbers of dimensions such as user, hashtag, location and time. To address this issue, we give axioms that any metric of suspiciousness should satisfy and propose an intuitive, principled metric that satisfies the axioms, and is fast to compute. We further propose CROSSSPOT [7] to spot suspicious regions and sort them in importance order. Empirical results show that CROSSSPOT improves the F1 score by 68% when capturing hashtag-hijacking and retweet-boosting in Weibo datasets spanning 0.3 billion posts.

Impact.

- CATCHSYNC [6] was selected as one of the best paper finalists of KDD 2014 with **37** citations. The algorithm has been taught in (1) CMU 15-826 "Multimedia Databases and Data Mining", (2) UMich EECS 598 "Graph Mining and Exploration at Scale", and (3) ASONAM 2016 Tutorial "Identifying Malicious Actors on Social Media" by Kumar et al. from UMCP.
- LOCKINFER [8] was the first paper to introduce the concept of *Camouflage* in fraud detection, and it now has **26** citations and it has appeared in KAIS 2015 [9]. CROSSSPOT [7] has **13** citations and it has appeared in TKDE 2016 [10]. The KDD 2016 best research paper by Hooi et al. cited all these three algorithms.
- Our survey paper about current trends and future directions on suspicious behavior detection has appeared in IEEE Intelligent Systems (IS) [11]. It was ranked as the top 10 most frequently downloaded documents in the IS from January to March 2016.

T2. Structuring behavioral content into heterogeneous information network

In Theme T1, I work on modeling social spatiotemporal contexts in behavioral data: the contexts are naturally well-structured and thus can be represented as multidimensional behavior networks. However, the behavioral content, especially the text corpus in tweets and papers, is information-rich but *unstructured*. Previously the text was represented as bag-of-words or topic models. Only by structuring the text into a rich attributed heterogeneous information network we will deeply understand the behavioral content. For example in Figure 6, can we automatically discover attribute names and values of different types of entities such as age of a person and prime minister of a country? Google's BIPERPEDIA [13] utilized their petabytes of users' fact-seeking queries (e.g., "canada prime minister") to harvest the attribute names. However, we are not able to access such query logs; moreover, we expect to extract the concrete values along with the names. We address this problem from a very different angle: leveraging *distant supervision* and "*data redundancy*" to discover *meta patterns* for automatic attribute discovery.

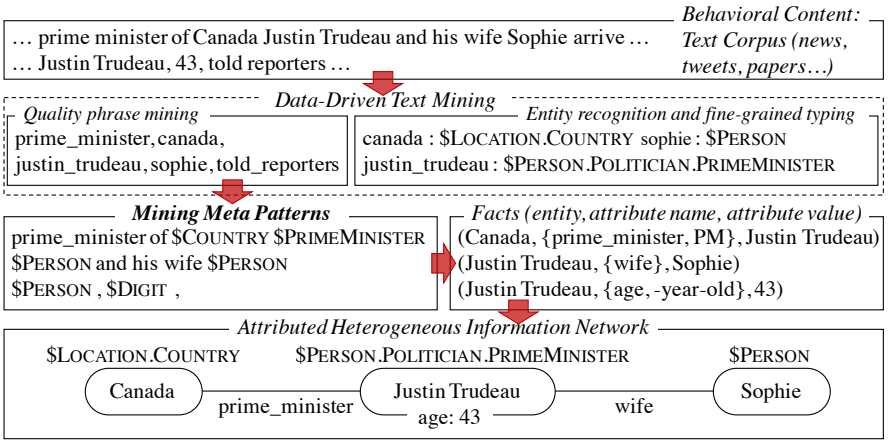


Figure 5: From text to information network: METAPAD [12] is a data-driven method that mines and uses meta patterns to discover attribute names and values of entities.

- Unlike weak supervision that relies on manually-specified entities, distant supervision (DS) aims at reducing expensive human labeling by utilizing information in Knowledge Bases (e.g., Freebase, Wikipedia). Recently, the UIUC data mining group has developed data-driven quality phrase mining SEGPHRASE [14] and entity recognition and typing CLUSTYPE [15] based on DS.
- Massive text corpus often provides sufficient data redundancy for us to derive informative frequent patterns. With the data-driven entity recognition and typing, the entity mentions can be replaced with their class types, and then attributes will appear frequently together with their corresponding entity classes. For example, by replacing "US", "Russia" and many other country names with "$COUNTRY", the meta pattern "president of $COUNTRY" becomes apparent in the text because this segment's frequency significantly deviates from what is expected from the frequencies of the symbol and words.

A **meta pattern** refers to a segment of class symbols (e.g., $COUNTRY, $PERSON), words (e.g., "wife"), phrases (e.g., "prime_minister") and possibly punctuation marks that appears contiguously and frequently in the text and serves as an integral semantic unit of the classes in certain context. We develop a framework called Meta Pattern-driven Attribute Discovery (METAPAD) [12] that integrates the text mining techniques and then mines quality meta-patterns of frequency, completeness, informativeness and appropriate granularity. We take the meta patterns such as "prime_minister of $COUNTRY $PRIMEMINISTER" and "$PERSON , $DIGIT ," back to the text to extract who is the prime minister and what is the age. Since we do not have the query data, we use BIPERPEDIA^{-q} (i.e., no query log) as the baseline taking every sentence of the corpus as "query" input. METAPAD improves the F1 score of detecting

attribute names by 48–190% from three news datasets and as much as 208% from tweets. We further extend it to PubMed corpus and successfully discover thousands of attributes of biomedical concepts. Moreover, this methodology is language independent.

Impact.

- METAPAD was collaborated with Dr. Talor Cassidy, Dr. Lance M. Kaplan and Dr. Timothy R. Hanratty from Army Research Lab (ARL). The package is transferring to ARL in the Network Science Collaborative Technology Alliance (NS CTA) project.

T3. Integrating behavior network and information network for behavior summarization

Fusing structured and unstructured data of human behaviors is substantial for in-depth behavioral analysis. Even bringing the quality phrases that were extracted from the unstructured content into behavior modeling has already been valuable and challenging. For example, given phrases and spatiotemporal contexts of tweets, can we automatically detect and summarize events from the multidimensional data? High-order tensor assumes that every behavior has exactly one value in every dimension, for example, a tweet has one user, one phrase and one hashtag. However, the real case is a behavior may have multiple values in a dimension. We propose "two-level matrix" that enables representing every single tweet: (1) the first/second level of columns are dimensions/dimensional values, (2) the first/second level of rows are time slices/behaviors. Thus, an event summary can be defined as a set of interesting dimensions and consecutive time slices, a set of interesting values in each selected dimension and a set of behaviors (tweets) in each selected slice, which forms a "Tartan" in the matrix. We develop a propagation method CATCHTARTAN [16] to capture the Tartans in a principled and scalable way: it determines the meaningfulness of every operation of updating the Tartan with the Minimum Description Length (MDL) principle. Empirical results show that it outperforms the tensor-based approaches, requires no parameter and provides comprehensive summaries of local events in tweets and research trends in academic data.

		User			Location		Hashtag		Phrase			
		u_1	u_2	u_3	l_1	l_2	h_1	h_2	p_1	p_2	p_3	p_4
t_1	i_1	1	0	0	1	0	0	1	0	1	1	0
	i_2	1	0	0	0	0	0	0	1	0	0	0
t_2	i_3	0	1	0	1	0	0	1	0	2	0	0
	...											
t_3	i_n	0	0	1	1	0	0	2	0	1	1	0
	...											

Florida *#PrayingForHaiti* *"please_donate"* *"those in need"*

Figure 6: Each row is a tweet. [16] detects and summaries events by looking for *Tartans* in "two-level matrix" which represents every dimension of tweets including time, location and *phrase*.

Fun facts.

- CATCHTARTAN [16] was accepted as full paper in KDD 2016 (70 from 784 submissions; 8.9%). It is the 1^{st} conference paper that Dr. Han (UIUC) and Dr. Faloutsos (CMU) co-authored, though they have been predicted to be co-authors for long.
- This is the $\underline{9}^{th}$ paper I have collaborated with Dr. Faloutsos as the 1^{st} author after my $\underline{9}$-month visit in CMU. Actually, Tartan is the team nickname of CMU. I watched lots of games, women's and men's, football and baseball, swimming and tennis.

II. FUTURE RESEARCH DIRECTIONS

When more real-world applications (e.g., economics, politics, business intelligence) are incorporating data science, the data-driven behavioral analytics should also be incorporating with other field sciences (e.g., behavioral science, psychological science, social science). The need for intelligence, trustworthiness and scalability in user-oriented systems will only increase.

Long-Term Vision: Interdisciplinary Research and Real-World Impact of Behavioral Analytics

The future of behavioral analytics locates at collaboration between data scientists and behavioral psychologists, which is a formidable combination that could not be accomplished without pooling these unique skills that differ by scientific training.

1. Bringing psychological expertise into data technology. Data in itself is nothing. We can only create value if we turn it into information. Data expertise alone is not sufficient to truly understand human behaviors. Behavioral scientists know more about the human brain. They are specialized in cognitive psychology. How do people perceive? How do they reach a decision? This kind of knowledge is crucial if we want to intervene users' behavior. Take the social contextual recommendation as an example. When I was struggling in understanding how Weibo users decide to forward or ignore a message, a Science paper [17] emerged to help. Salganick et al. adopted an experimental approach to the study of social influence in cultural markets, inspired by which I reconsidered the information adoption mechanism of Weibo and proposed that preference and influence were two major contextual factors. Experiments on real large data demonstrated my assumption (see Figure 2) and the proposed model CONTEXTMF significantly improved Weibo's recommender systems. It is important to realize that vision if we bring real psychological expertise into our data science.

2. Combining psychological discoveries with data science. Data-driven approaches have enormous potential to change the way psychological scientists observe human behavior. Big data leads researchers to a point where they can collect behavioral information without sampling human participants at all. For example, social media, e.g., Facebook, Twitter, are the new "macroscopes of human behavior". Technology such as smart phones and wearable sensors can gather information on physical activity, social interactions, and so on. This offers us a clue to understanding basic psychological principles and behavioral theories for experimentation. Psychological scientists and data scientists have a vested interest in working together to yield more explanatory insight that can change the world.

3. Deploying scalable behavior modeling for real systems. To scale complex behavior models to the high volume of data, my research efficiently exploits the structure of multidimensional behavior networks and heterogeneous information networks. The ultimate goal is to deploy evolutionary analysis (e.g., [5, 4]) and scalable algorithms (e.g., [6, 16]) across a wide variety of applications.

Mid-Term Research Plan

With respect to my shorter term plan (first 3-5 years), I provide a more detailed outline of the thrusts and challenges that I am planning to tackle, both contributing to the data-driven framework of behavioral analysis that support decision-making processes.

D1. Intelligence: Integrating networks for in-depth behavioral analysis

Integrating unstructured data for accurate and interpretable contextual behavior modeling. In my thesis and postdoctoral research, I have improved the state-of-the-art in modeling user behavior with structured data: predicting behaviors with contexts and catching suspicious behaviors. However, if we don't model the content with structures, big data means big mess. Fortunately, the data-driven meta pattern mining approaches shed light on automatically structuring the content into a rich network of entities and attributes. How to grab this opportunity to develop accurate and interpretable predictive models needs to be investigated.

Predicting behavioral contexts over predicting contextual behaviors. Suppose we are able to predict with 100% accuracy if a behavior will happen given full contexts of it. Then given a subset of the contexts, can we predict the set of the rest contexts? For example, if one wants to solve the fraud detection problem on Weibo, which experts, papers and algorithms he/she should find, read and try? Assume that we have represented the applications, datasets and other entities as nodes in an n-node heterogeneous network, the complexity of searching for the optimum is too high: the space of solutions is $O(n!)$. How to reduce the complexity into a practical level by pruning impossible permutations with insights from behavioral and psychological sciences is interesting and challenging.

D2. Trustworthiness: Structuring reliable behavioral content

Our meta-pattern mining is a general, data-driven methodology to be used across multiple NLP tasks such as entity recognition and (fine-grained) typing in an automated way: the attributes are discovered by looking for meta patterns after aggregating typing results, and further the typing module can be improved by correcting the types of entities with features of their attributes. However, the meta-pattern mining still generates some low quality patterns and wrong attribute values (e.g., the text "McDonald's U.S. President..." may mistake a McDonald executive as a U.S. president). In order to structure reliable information from behavioral content, it is necessary to enhance mining results with majority voting-based conflict resolution, sentence structure-based entity refinement, conditional functional dependency rule mining, and advanced truth discovery from massive documents.

Contributing to the community. I am focused on studying and promoting Data-Driven Behavioral Analytics. I have given two 3-hour tutorials in major conferences (i.e., ICDM 2015 [18] and CIKM 2016 [19]). Each had 50+ audience, and we won the honorium for both. I also have written two book chapters about user behavior modeling [20, 21].

Funding opportunities. First, based on the idea of CATCHTARTAN [16] (submitted in Feb 2016), I wrote the 2^{nd} section taking over 8 from 15 pages of the proposal "NSF III: Small: Multi-Dimensional Structuring, Summarizing and Mining of Social Media Data" in Oct 2015. It has been awarded to the PI Dr. Jiawei Han by NSF IIS (08-01-2016 to 07-31-2019, $500,000). I am the only major supported member[4]. Besides "Information Integration and Informatics" (III), "Secure and Trustworthy Cyberspace" (SaTC) also encourages the study of behavior analysis. Second, I have extensively collaborated with and visited the Army Research Lab (ALC and APG) in the NS CTA I1 and I4 projects. I significantly contributed in writing the whitepaper for Y8/9 (2016–2018). Third, I have international relations with China's institutes (e.g., Tsinghua, Microsoft Research Asia) and IT companies (e.g., Tencent, Alibaba).

REFERENCES

[1] **Meng Jiang**, Peng Cui, Rui Liu, Qiang Yang, Fei Wang, Wenwu Zhu, and Shiqiang Yang. Social contextual recommendation. In *ACM CIKM*, 2012.

[2] **Meng Jiang**, Peng Cui, Fei Wang, Qiang Yang, Wenwu Zhu, and Shiqiang Yang. Social recommendation across multiple relational domains. In *ACM CIKM*, 2012.

[3] **Meng Jiang**, Peng Cui, Nicholas Jing Yuan, Xing Xie, and Shiqiang Yang. Little is much: Bridging cross-platform behaviors through overlapped crowds. In *AAAI*, 2016.

[4] **Meng Jiang**, Peng Cui, Fei Wang, Xinran Xu, Wenwu Zhu, and Shiqiang Yang. Fema: Flexible evolutionary multi-faceted analysis for dynamic behavioral pattern discovery. In *ACM SIGKDD*, 2014.

[5] **Meng Jiang**, Peng Cui, Fei Wang, Wenwu Zhu, and Shiqiang Yang. Scalable recommendation with social contextual information. *IEEE TKDE*, 2014.

[6] **Meng Jiang**, Peng Cui, Alex Beutel, Christos Faloutsos, and Shiqiang Yang. Catchsync: Catching synchronized behavior in large directed graphs. In *ACM SIGKDD (**best paper finalist**)*, 2014.

[7] **Meng Jiang**, Alex Beutel, Peng Cui, Bryan Hooi, Shiqiang Yang, and Christos Faloutsos. A general suspiciousness metric for dense blocks in multimodal data. In *IEEE ICDM*, 2015.

[8] **Meng Jiang**, Peng Cui, Alex Beutel, Christos Faloutsos, and Shiqiang Yang. Inferring strange behavior from connectivity pattern in social networks. In *PAKDD*, 2014.

[9] **Meng Jiang**, Peng Cui, Alex Beutel, Christos Faloutsos, and Shiqiang Yang. Inferring lockstep behavior from connectivity pattern in large graphs. *KAIS*, 2015.

[10] **Meng Jiang**, Alex Beutel, Peng Cui, Bryan Hooi, Shiqiang Yang, and Christos Faloutsos. Spotting suspicious behaviors in multimodal data: A general metric and algortihms. *IEEE TKDE*, 2016.

[11] **Meng Jiang**, Peng Cui, and Faloutsos Christos. Suspicious behavior detection: Current trends and future directions. *IEEE Intelligent Systems*, 2016.

[12] **Meng Jiang**, Jingbo Shang, Taylor Cassidy, Lance Kaplan, Timothy Hanratty, and Jiawei Han. Metapad: Meta pattern-driven attribute discovery in massive text corpora. In *ACM WSDM (in peer review)*, 2017.

[13] Rahul Gupta, Alon Halevy, Xuezhi Wang, Steven Euijong Whang, and Fei Wu. Biperpedia: An ontology for search applications. *VLDB*, 2014.

[14] Jialu Liu, Jingbo Shang, Chi Wang, Xiang Ren, and Jiawei Han. Mining quality phrases from massive text corpora. In *ACM SIGMOD*, 2015.

[15] Xiang Ren, Ahmed El-Kishky, Chi Wang, Fangbo Tao, Clare R Voss, and Jiawei Han. Clustype: Effective entity recognition and typing by relation phrase-based clustering. In *ACM SIGKDD*, 2015.

[16] **Meng Jiang**, Christos Faloutsos, and Jiawei Han. Catchtartan: Representing and summarizing dynamic multicontextual behaviors. In *ACM SIGKDD*, 2016.

[17] Matthew J Salganik, Peter Sheridan Dodds, and Duncan J Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *science*, 311(5762):854–856, 2006.

[18] **Meng Jiang** and Peng Cui. Behavior modeling in social networks: From micro to macro. *IEEE ICDM (tutorial)*, 2015.

[19] **Meng Jiang**, Peng Cui, and Jiawei Han. Data-driven behavioral analytics: Observations, representations and models. *ACM CIKM (tutorial)*, 2016.

[20] **Meng Jiang** and Peng Cui. Mining user behaviors in large social networks. *Big Data in Complex and Social Networks*, 2016.

[21] **Meng Jiang**. Behavior modeling in social networks. *Encyclopedia of Social Network Analysis and Mining, 2nd edition*, 2016.

[4]NSF IIS 16-18481. `http://hanj.cs.illinois.edu/projs/social\_media.htm`