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 the *interaction* of individuals with themselves and with their environment.[1] Thanks to Information Technologies, online human behaviors are broadly recorded at an unprecedented level. This gives us an opportunity for getting insights into behaviors and our societies from large-scale real data. The Department of Defense (DoD) listed *Computational Modeling of Human Behavior* as one of the high-priority topics in DoD Basic Research. In order to provide a fundamental understanding and predictive capability of behavior dynamics from individuals to societies, behavioral analysis has to face four complexities: (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 (e.g., suspicious purposes) are a nontrivial part of behavior modeling; and (4) user preferences and sentiments on *unstructured content* are important driving factors. I focus on analyzing behavioral data with network models and algorithms 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 accuracy and interpretability.

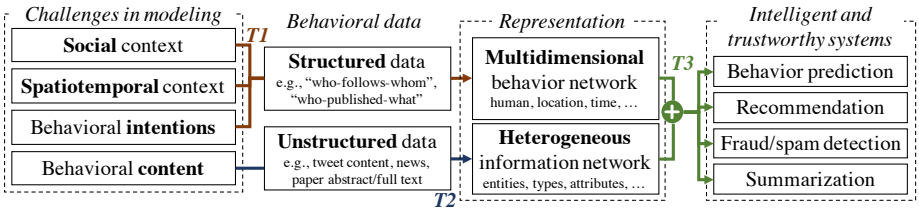

Figure 1: Data-driven behavioral analytics towards intelligence and trustworthiness.

I. THESIS AND POSTDOCTORAL WORK

My thesis titled "Modeling Complex Behaviors in Social Media" focuses on T1, and my 2-year postdoctoral research is broken down into T2 and T3. During my 5-year Ph.D., I was fortunate to join the long-term collaboration between Tencent Weibo and Tsinghua, which put me in the unique position of having access to massive real data generated by millions of users.

T1. Mining behavior network with social spatiotemporal contexts

T1.1 Modeling social spatiotemporal contexts for behavior prediction

Weibo suffers from *low* conversion rate: Users received too much information and generated fewer than six retweets for every 100 feed requests. Can we recommend tweets by predicting their behaviors to address the issue of information overload? We observed from the real data that personal preference on posted content and interpersonal influence from the poster are two significant factors that determine users' decisions. We proposed 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. Empirical results on the Weibo dataset demonstrate that CONTEXTMF reduces the root-mean-square error (RMSE) of prediction by 21%. Furthermore, I proposed HYBRIDRW [2] for *cross-domain behavior modeling* to "cold-start" recommendation, XPTRANS [3] for *cross-platform behavior modeling*, and FEMA [4] for *spatiotemporal behavior modeling*. CONTEXTMF [1] and HYBRIDRW [2] are the 3^{rd} and 9^{th} most cited papers of CIKM 2012 with **122** and **47** citations. CONTEXTMF+ [5] 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 given certain accounts many additional followers, making them seem more legitimate or famous. Essentially, we consider the detection problem by revealing the fraudsters' manipulation on the network that they have to do. As shown in Figure 2, the spikes on the out-degree distribution indicate millions of anomalous nodes on the network. The fraudsters exhibit behavior that is (1) synchronized (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 developed a scalable and parameter-free algorithm, CATCHSYNC [6], to quantify the two concepts. For a follower, the *synchronicity* score measures how the followees are similar with each other, and the *normality* score measures how the followees are similar with other nodes in the network. We proved that given the normality score, the synchronicity score has a parabolic lower bound. 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. CATCHSYNC [6] was selected as one of the best paper finalists of KDD 2014 and currently has **40** citations. It has been taught in CMU 15-826 and UMich EECS 598. Furthermore, I proposed a principled metric to evaluate the suspiciousness in multi-dimensional behavioral data, and a scalable algorithm, CROSSSPOT [7, 8], to detect the hashtag-hijacking and retweet-boosting in Weibo.

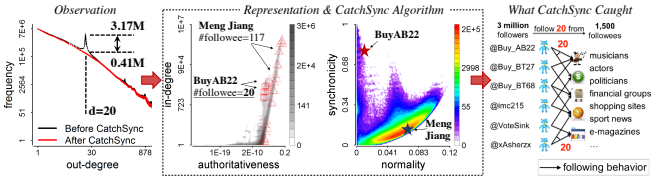

Figure 2: CATCHSYNC [6] spots that fake followers consistently connect to customers of similar features: Their synchronized behaviors create spikes on degree distributions.

[1] Behavior – Wikipedia, the free encyclopedia: https://en.wikipedia.org/wiki/Behavior

T2. Structuring behavioral content into heterogeneous information network

The contexts are naturally well-structured, however, the behavioral content, especially the text corpus in tweets and papers, is information-rich but *unstructured*. Structuring the text into heterogeneous information networks enables deep understanding of the behavioral content. Given the text data, e.g., a news corpus with sentence "President Blaise Compaoré's government of Burkina Faso was...", can we extract (1) the ⟨entity type, attribute name⟩ pair, e.g., ⟨$COUNTRY, president⟩ and (2) the ⟨entity, attribute name, attribute value⟩ tuple, e.g., ⟨Burkina Faso, president, Blaise Compaoré⟩? The state-of-the-art open IE systems learn syntactic and lexical patterns of expressing relationships for extracting entity-level tuples. However, they may generate incorrect or imprecise extractions (e.g., ⟨President Blaise Compaoré, have, government of Burkina Faso⟩ in which "have" is not a good attribute name for the entity by ignoring the type information). Our idea is to investigate joint extraction of the type-level ⟨entity type, attribute name, attribute value type⟩ tuples (adding "value type" into the pair) and the entity-level tuples because these two extraction processes can mutually enhance each other. We propose a novel methodology, *Meta Pattern Mining*, that mines frequent and informative semantic patterns indicating type-level tuples, called *meta patterns*. If we replace "Burkina Faso" with $COUNTRY and replace "Blaise Compaoré" with $POLITICIAN, and carefully segment the sentence, we can generate meta pattern "president $POLITICIAN's government of $COUNTRY". We develop a Meta Pattern-driven Attribute Discovery (METAPAD) framework [9], which first pre-processes text data to extract entities and their types as input in a data-driven and distantly-supervised manner and then mines the meta patterns for attribute discovery. Experiments on news, tweets and biomedical text data demonstrate that METAPAD improves the F1 scores over the state-of-the-art by 32–51% in ⟨entity type, attribute name⟩ extraction and by 26–35% in ⟨entity, attribute name, attribute value⟩ extraction. This is a collaboration with Dr. Taylor Cassidy and Dr. Lance M. Kaplan from U.S. Army Research Lab.

T3. Integrating behavior network and information network for behavior summarization

Fusing structured and unstructured human behavioral data is imperative 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 methods assume that every behavior has exactly one value in every dimension; for example, a tweet has one user, one phrase, and one hashtag. However, in real cases, a behavior may have multiple values in a given dimension. We proposed CATCHTARTAN [10] that represents the behavioral data with a hierarchical structure and uses a greedy algorithm based on the Minimum Description Length principle to find the summaries. Empirical results show that it outperforms the tensor-based approaches, requires no parameters and provides comprehensive summaries of local events in tweets and research trends in academic data. Based on this work, I contributed significantly to the technical content in the proposal "NSF III: Small: Multi-Dimensional Structuring, Summarizing and Mining of Social Media Data." It has been awarded to the PI, Professor Jiawei Han.

II. FUTURE RESEARCH DIRECTIONS

With respect to my mid-term plan (first 3-5 years), I have provided directions and challenges that I am planning to tackle, both contributing to data-driven behavioral analysis that supports decision-making processes. (1) Integrating structured and unstructured data for *intelligent* behavioral analysis. If we don't model the content with structures, big data can turn into a big mess. Fortunately, the data-driven meta-pattern mining approaches facilitate automatically structuring the content. I will leverage this opportunity to develop accurate and interpretable predictive models. (2) Structuring *trustworthy* information networks from behavioral content. In order to structure reliable information, it is necessary to enhance mining results with majority voting-based conflict resolution, sentence structure-based entity refinement, and conditional functional dependency rule mining.

The future of behavioral analytics lies in the intersection of data science and behavioral psychology, which is a formidable combination wherein each group brings to the table unique skills that differ by scientific training. I will conduct cross-disciplinary research and build real-world impact of behavioral analytics in the following directions: (1) Bringing psychological expertise into data technology. Behavioral scientists, who typically specialize in cognitive psychology, know more about the human brain. This kind of knowledge is crucial if we want to interpret users' behavior. (2) Facilitating psychological discovery processes with data science. (3) Deploying scalable behavior modeling for real systems. The direct impact of my work on Weibo's revenue stream is compelling evidence for the value of scalable algorithms across a wide variety of applications.

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**, 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.

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

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