Research Statement

Feida ZHU
School of Information Systems, Singapore Management University
Tel: (65) 6808-5101; Email: fdzhu@smu.edu.sg
05 (Day) 01 (Month) 2024 (Year)

Background and Mission

Our time has been characterized by an explosion of data of all sorts. In particular, the recent blossoming of social network services has provided everyone with an unprecedented level of ease of access and fun with sharing information of all kinds. Public social data reveal a surprisingly large amount of information about an individual that is otherwise unavailable. The business, consumer and social insights attainable from the big and dynamic social data available today are critically important and immensely valuable in a wide range of applications for both the private and public sectors.

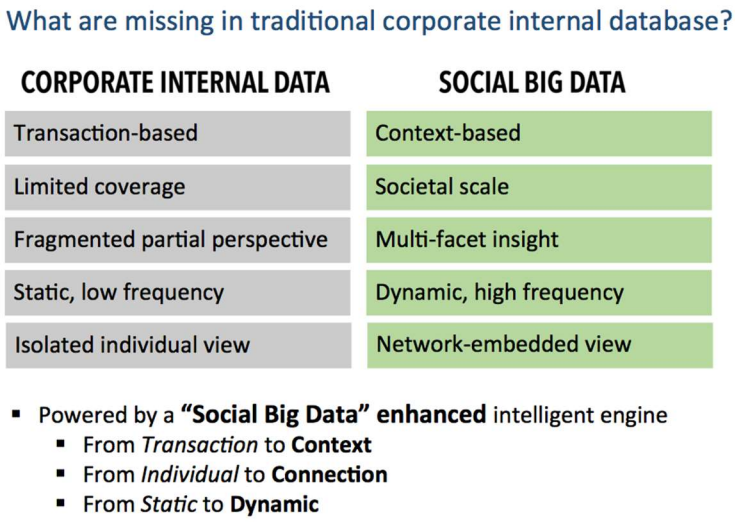
On the other hand, the core of all social and business activities is "people". An understanding of people and their lives at societal-scale yet with individual-granularity is the key to almost all real-life applications. It is therefore my research mission -- *to spearhead a data-driven technology-enabled business and social intelligence paradigm by focusing on data of the people, by the people, and for the people.*

Data of the people refers to the nature of our data -- everything describing individuals and their lives -- and our consistent approach of integrating and organizing all our data with the thread of every individual. *Data by the people* highlights our focus on user-generated content and user behavior data for user profiling in our analysis and mining. *Data for the people* indicates that the goasl and principles of our research are to bring actionable intelligence out of data for the ultimate benefit of people.

Central to my research mission are two research dimensions of equal importance which complement and inform each other: *user profiling* and *event profiling*.

For user profiling, I aim at knowledge from enrichment at societal scale and individual granularity simultaneously.

For event profiling, I strive for agility from real-time response to take advantage of the instant nature of multi-source heterogeneous social data.



These things complement each other. While user profiling aims to best characterize each individual and identify the "right person" in business settings, event profiling aims to describe how individuals interact with one another, provide context to user profiling and capture the "right timing" on the other hand.

By integrating the user and event dimensions, we are able to build real data-powered systems to answer interesting questions. What can we tell from the social data on the context of consumer

behaviour, such that it is possible to enrich the transaction-based data of traditional corporate databases? How can the power of social connections be unleashed to identify potential high-value customers and perform cost-effective risk management? How to achieve dynamic social listening on 200 million users and detect marketing opportunities in real-time based on bursty events?

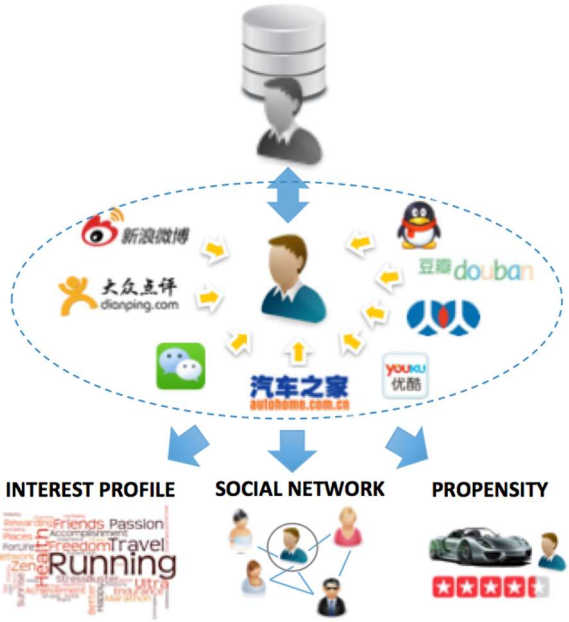
Research Map

My research goal is to achieve actionable intelligence by exploring and focusing on data of the people, by the people and for the people. In particular, we leverage primarily the abundant social data and aim to build a real-time cross-platform social data mining system to achieve and integrate two complementary dimensions: (I) comprehensive and accurate user profiling down to individual level; and (II) responsive and intelligent event profiling at societal scale for bursty and viral topics.

The Two Dimensions of "Social Data" Mining

(1) <u>User Profiling -- Knowledge From Enrichment</u>

In my studies, *users* refers to real-life persons, instead of online accounts (one user could have several accounts on different social platforms). My research focus here has been motivated by two critical issues in achieving an in-depth and all-around understanding of individual user for real-life business applications.



First, since the information of a user from the current social scene is fragmented, inconsistent and disruptive, the key to unleashing the true power of social media analysis is to link up all the data of the same user across different social platforms; Secondly, one important question in social data mining which has not been well explored is --- "How would all the analytical results about the *online* social data impact our *offline* real life?" For example, all the research findings on user influence would remain inconsequential if we are not able to establish the linkage between the online and offline world.

As shown in the figure, the value of user profiling based on cross-platform user identity linkage and online-offline cross-mining is that we would be able to significantly enrich a business's knowledge about their customer with three kinds of information that are usually unavailable in typical corporate database: all-around user interest profile, a user's real-life core social network and a user's propensity model for various products and services.

1. <u>Cross-platform User Identity Linkage</u>

   This problem is of huge practical value and research difficulty at the same time due to a number of challenges including data misalignment in the forms of platform difference, behavior asynchrony, data veracity and imbalance. My work has collectively attacked the problem from several frontiers.

   My HYDRA framework in SIGMOD'14 and TKDE'15 article with my research colleagues has been the first to comprehensively and systematically consider all user data types available

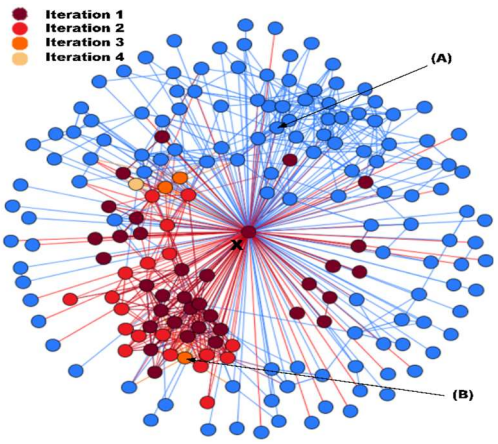
on social platforms and integrate them into a unified linking framework. In particular, the technical breakthrough came from taking advantage of two important features unique to social data: user behavior trajectory along the temporal dimension, and the users' core social network structure. Experiments on real data of millions of users demonstrated the effectiveness of framework.

The ICDM'15a work provides an unsupervised solution to handle the absence of ground-truth data typical in these problem settings. The KDD'16 work extends beyond a two-platform user linkage setting and presents a latent user space model to give a more natural and robust solution to the multi-platform user linkage problem. As user features are usually avoidable in such scenarios, how to assign weights to various features has remained a challenging issue. My collaborative DASFAA'16 work, which has won a **Best Paper Award**, provides a systematic data-driven framework to consider feature discriminative scores to make the best use of user peculiarity as exhibited on social platforms.

2. Online-Offline Cross-Mining

While online social media research has been flourishing, much less progress has been made in crossing the online-offline boundary and exploring users' offline real-world information based on what we can observe from their online data. My research agenda here is to fill this gap and establish the connection.

Some of my best work was in WebSci'12, which presented an effective algorithm to discover users offline real-world social connections by purely examining their online Twitter follower network alone. This is the first time users' offline core social connections were identified from their online social network based on three proposed principles: *mutual reachability, friendship retainability* and *community affinity*.



This work provides a new foundation for many exciting applications and future works including robust user modeling, business competitive analysis, user profile matching, spammer detection, etc. Based on this work, the next work in DASFAA'13 was to propagate dynamically user attribute labels in the relationship network. The corresponding demo system won the **Best Demo Award (Runner-Up**) at DASFAA'13. Furthermore, the ApWeb'16 paper more deeply explored the data to identify users' intimate relationships, which has important application in risk management for the financial industry.

To gain a deeper understanding of the interplay between users' online and offline relationship and behavioral patterns, the network formation principles between offline and online friends and friendship influence were examined in NetSciX'16a.

Another interesting line of work that goes from the online to the offline settings sought to find out who were the 'lurkers' in an online social network: the silent majority who publish little but passively receive information. The series of work that has appeared in ICWSM'15,'16 and NetSciX'16b has been pioneering in its effort to systematically define lurkers, and propose algorithms to find them and characterize their behavior.

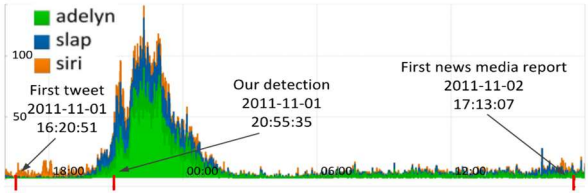
Online social data can also provide clues to users' offline personal credit information, as shown in my joint PAKDD'16 work. It covers how we extract users' profile, content, behavior

and network information from their social data to build an ensemble learning model which can predict a user's personal credit capacity to an accuracy of 70%.

(2) Event Profiling --- Agility from Real-time Responsiveness

Perhaps the most important and unique feature of social media compared against all other traditional news media is the real-time responsiveness of the data. For example, it has been observed that, in life-critical disasters of societal scale, Twitter is the most important and timely source from which people find out and track the breaking news before any mainstream media picks up on it and rebroadcasts the footage. Consequently, it is essential that we are able to conduct mining and analysis for the huge-volume data flow in a real-time fashion. My work in event profiling covers the full spectrum from detection to visualization as follows.

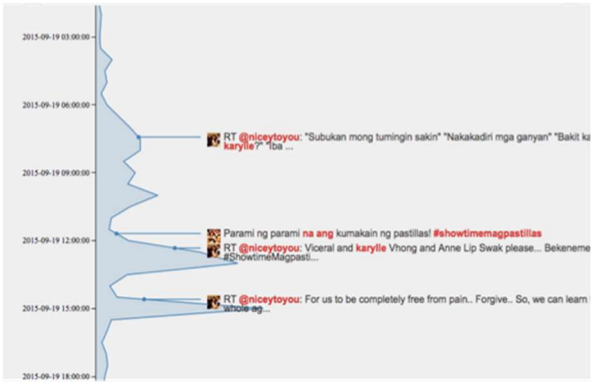
a) Bursty Event Detection: Identify bursty and viral events as soon as they emerge in real-time. Bursty topics are events which trigger a surge of population-wise attention. The joint work in ACL'12 proposed the first algorithm to find such topics from Twitter in an offline fashion.



To achieve real-time responsiveness, my collaborative ICDM'13 work proposed a novel mining framework called "TopicSketch" which was the first work that achieved real-time detection on social media for bursty topics with no pre-defined keywords. The algorithm has been further improved in TKDE'16a to achieve better robustness and efficiency. Once a bursty event is detected, it is also important yet nontrivial to trace back to the very source node of the information diffusion, which is addressed in ICDM'15b.

b) Event Virality Prediction: Predict event popularity and set up dynamic monitoring mechanisms at an early stage. Not all bursty topics are worth following, which leads to a challenging question of how to predict the virality of a topic at the early stage of its diffusion. The BigData'15 paper on "Modelling Cascades Over Time in Microblogs" provides accurate early prediction, and combined with its bursty topic detection algorithm, provides an effective means to capture bursty and viral topics right before they trigger momentum and go viral.

c) Event Summarization and Contextualization: Summarize events to construct a semantically coherent and easy-to-understand story-line in temporal order. Provide event context by integrating related content from different sources to enrich event understanding. Our PKDD'15 and TKDE'16b work on online summarization of dynamic network diffusion, and ApWeb'16 paper, and WAIM'16's **Best Demo Paper** all address the problem of how to summarize not just the textual aspect of detected events but to integrate the dynamic diffusion process and the user dimension in social events.



d) Event Visualization and Interaction: Visualize events with intuitive and informative data visualization techniques for easy understanding, and user-friendly features for convenient and fun exploration of the events. We have been integrating our research results on

various aspects of social events into a unified system called "CLEar" which stands for "*Cl*airvoyant *Ear".* A first version of the system has won the **Best Demo Runner-up Award** at VLDB'14.

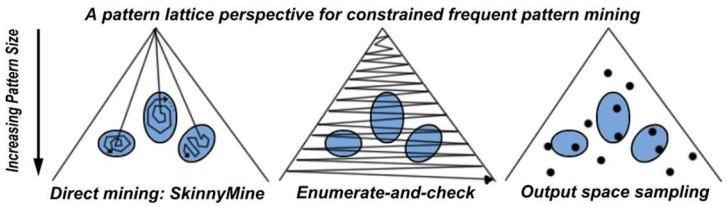
Two Computational Challenges of "Social Data" Mining

While pushing the frontiers along the two dimensions of users and events in social data mining, my research group's approach has been exploiting two important features unique to social data: a user's core social network structure and a user's behavior trajectory along the temporal dimension. This is a *network-aware* and *behavior-driven* approach.

1) <u>Efficient Network Structure Mining</u>

It is well-known that frequent pattern mining in graph setting is notoriously hard, especially in face of today's network scale. Most work on graph mining has been largely focused on graph transaction settings, where the input data is a large collection of small graphs. However, all the social network applications today present us with large single graphs. It has been shown that frequent pattern mining in single network settings is a much more challenging problem than its counterpart in transaction settings due to the existence of overlapping embeddings and much trickier support computation. My work on network pattern mining solve some of the large pattern mining challenges in the large single-graph setting.

My VLDB'11 paper on "Mining top-k large structural patterns in massive networks" was first to proposed how to find large patterns in massive graph data. Based on a novel concept called "r-spider", this work provides users for the first time with the capacity to reach and study the largest frequent patterns in big graph data in a reasonable amount of time.



With the boom in mobile social data and research on information diffusion, another kind of constrained pattern --- the "skinny" pattern, which is a graph pattern with a long back-bone from which short twigs branch out is important. There are important applications for the descriptive power of its long backbone to represent spatial and temporal trajectories in heterogeneous information networks, and of the short twigs the various kinds of associated information. My work in SIGMOD'13 proposed a new direct mining paradigm for efficient constrained frequent graph mining. Frequent patterns with certain structural constraints can be generated directly with minimum redundancy. This was impossible with traditional mining methodology in which patterns are grown in the order of increasing size.

2) <u>Behavior-driven Heterogeneous Data Mining</u>

Another characterizing feature of social media data --- user behavior --- is a key connecting element to integrate, understand and leverage the highly heterogeneous social data to go beyond what could otherwise be accomplished on the structured and unstructured data. In particular we pushed the user behavior element into the following mining tasks.

(1) Behavior-driven Topic Modeling

I proposed a B-LDA model with my research colleagues to incorporate user behavior into the LDA topic modeling in SDM'13. This is to better capture the user interactions which are critically important for topic analysis, user clustering and followee recommendation on social micro-blogging services such as Twitter.

(2) Behavior-driven Anomaly Detection

My joint work also used group-level user behavior, in SDM'12 and CIKM'12, to characterize anomaly collections, and identified spammer groups that are hard to catch with traditional point anomaly frameworks. We also used collective user rating behavior to model anomalous users and products in online review settings and proposed a unifying framework based on mutual dependency principles in ICDM'12. Extensions of these pieces of work have been published in DMKD'15.

(3) Behavior-driven Relationship Mining

We also studied the user follow links in Twitter network and developed in WebSci'12 a novel algorithm which, based on this piece of information alone, is able to identify with high accuracy those offline real-life friends of the target user. This work has profound potential impact. We also studied user-follower linkages to dynamically propagate user attribute-relationship labels with user input in a paper in DASFAA'13. In another work published at SocInfo'13, my research team re-visited the user-ranking problem for social networks and examined the problem from the user interaction perspective. A new angle to the problem based on interplay between "information" and "interaction" was offered.

Research Impact

The impacts of my research results along these lines have demonstrated themselves through over 80 peer-reviewed publications and two Best Paper Awards at top international venues, as well as over 30 invited talks, distinguished lectures and tutorials at major conferences and industrial events such as KDD'14, DASFAA'13, ICDE'15, WAIM'15 and DSC'16, and recognized professional services in the communities including Area Chair of ICDM'15 and '16, Senior PC of CIKM'16, Workshop Chairs of KDD'13 and '14, and PC member of over 50 top conferences.

Yet more importantly, the research findings and results have led to two major research labs I have jointly founded with industrial giants: (I) The Pinnacle Lab of Analytics with China Ping An Insurance Group and (II) The DBS-SMU Life Analytics Lab. Both are million-dollar industrial investments in multiple-year collaboration contracts. As the founding director of both labs, I have been consistently and productively driving academia-industry collaboration in a wide range of topics. They include activities with both research and practical value focusing on the theme of *data-driven technology-enabled business intelligence and innovation*. The mutually-benefit, academia-industry close interactions have resulted in an integrated system called "Pinnacle." It emphasizes the application of an intelligent mining engine with the following three components:

1) Data: an ever-increasing integrated database of close to 200 million Chinese users, and event profiles of societal-scale yet down-to-the-individual level, which has contributed to various research projects and teaching materials forthe SMU community and beyond.

2) Algorithms: a suite of effective and scalable algorithms that underlie the real-time cross-platform social data mining and analysis;

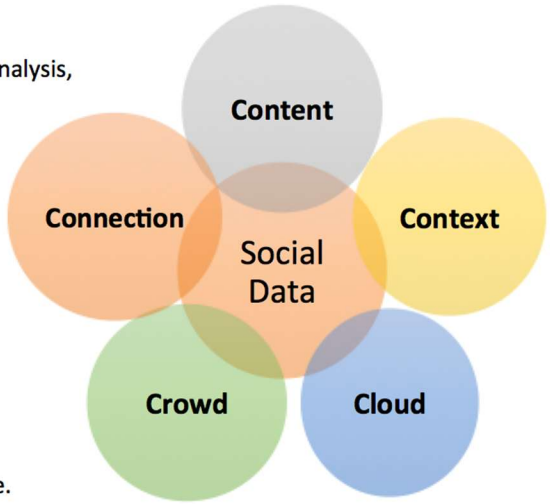
3) System: an online system to perform multi-dimensional networked analysis of user profiles and real-time viral topics and emerging trends, offering user-friendly interaction and discovery-driven exploration for a range of business applications including targeted marketing, risk management, customer acquisition and personalized CRM.

Due to the success of these collaborations, the various applications and platforms powered by my research from these two labs have gained global recognition and visibility as 'best practice' and pioneering efforts in exploring and applying the power of social data to drive business innovation and intelligence.

Future Work

My research agenda in the near future will be focused on the theme of analyzing and mining data of the people, by the people and for the people along, in particular, the dimensions of five "C" unique to social data -- *C*ontent, *C*ontext, *C*onnection, *C*rowd and *C*loud.

- **Content**
  - Profile categorization, topic modeling, sentiment analysis, interest mining, etc.
- **Context**
  - Location, temporal analysis, behavior trajectory, community, etc.
- **Connection**
  - Relationship mining, core network detection.
- **Crowd**
  - Tap into the power of collective human wisdom.
- **Cloud**
  - Big data needs a different underlying infrastructure.



Reference

- [VLDB'11] "*Mining top-k large structural patterns in a massive network*", F Zhu, Q Qu, D Lo, X Yan, J Han, PS Yu
- [SDM'12] "*Detecting extreme rank anomalous collections*", H Dai, F Zhu, EP Lim, H Hwa Pang
- [CIKM'12] "*Mining coherent anomaly collections on web data*", H Dai, F Zhu, EP Lim, HH Pang
- [WebSci'12] "*When a friend in twitter is a friend in life*", W Xie, C Li, F Zhu, EP Lim, X Gong
- [ACL'12] "*Finding bursty topics from microblogs*", Q Diao, J Jiang, F Zhu, EP Lim
- [ICDM'12] "*Detecting anomalies in bipartite graphs with mutual dependency principles*", H Dai, F Zhu, EP Lim, HH Pang
- [SDM'13] "*It is not just what we say, but how we say them: Lda-based behavior-topic model*", M Qiu, F Zhu, J Jiang
- [DASFAA'13] "*Dynamic label propagation in social networks*", J Du, F Zhu, EP Lim
- [SocInfo'13] "I*nformation vs interaction: An alternative user ranking model for social networks*", W Xie, AP Hoang, F Zhu, EP Lim
- [ICDM'13] "*TopicSketch: Real-time Bursty Topic Detection from Twitter*", W Xie, F Zhu, J Jiang, EP Lim, K Wang
- [DASFAA'13 demo] "*Twicube: A real-time twitter off-line community analysis tool*", J Du, W Xie, C Li, F Zhu, EP Lim
- [SIGMOD'13] "*A direct mining approach to efficient constrained graph pattern discovery*", F Zhu, Z Zhang, Q Qu
- [SIGMOD'14] "*Hydra: Large-scale social identity linkage via heterogeneous behavior modeling*", S Liu, S Wang, F Zhu, J Zhang, R Krishnan
- [VLDB'14] "*CLEar: a real-time online observatory for bursty and viral events*", R Xie, F Zhu, H Ma, W Xie, C Lin
- [PKDD'14] "*Interestingness-driven diffusion process summarization in dynamic networks*", Q Qu, S Liu, CS Jensen, F Zhu, C Faloutsos

- [TKDE'15] "*Structured learning from heterogeneous behavior for social identity linkage*", S Liu, S Wang, F Zhu
- [ICDM'15a] "*CNL: collective network linkage across heterogeneous social platforms*", M Gao, EP Lim, D Lo, F Zhu, PK Prasetyo, A Zhou
- [ICDM'15b] "*Information Source Detection via Maximum A Posteriori Estimation*", B Chang, F Zhu, E Chen, Q Liu
- [ICWSM'15] "*Characterizing Silent Users in Social Media Communities*.", W Gong, EP Lim, F Zhu
- [DMKD'15] "*Detecting anomaly collections using extreme feature ranks*", H Dai, F Zhu, EP Lim, HH Pang
- [BigData'15] "*Modelling cascades over time in microblogs*", W Xie, F Zhu, S Liu, K Wang
- [PAKDD'16] "*Personal credit profiling via latent user behavior dimensions on social media*", G Guo, F Zhu, E Chen, L Wu, Q Liu, Y Liu, M Qiu
- [ICWSM'16] "*On Unravelling Opinions of Issue Specific-Silent Users in Social Media*", W Gong, EP Lim, F Zhu, PHXU CHER
- [KDD'16] "*User Identity Linkage by Latent User Space Modelling*", X Mu, F Zhu, EP Lim, J Xiao, J Wang, ZH Zhou
- [DASFAA'16] "*When Peculiarity Makes a Difference: Object Characterisation in Heterogeneous Information Networks*", W Chen, F Zhu, L Zhao, X Zhou
- [ApWeb'16] "*When a Friend Online is More Than a Friend in Life: Intimate Relationship Prediction in Microblogs*", Y Lan, M Zhang, F Zhu, J Jiang, EP Lim
- [NetSciX'16a] "*A Comparison of Fundamental Network Formation Principles Between Offline and Online Friends on Twitter*", F Natali, F Zhu
- [NetSciX'16b] "*Posting Topics\ ne Reading Topics: On Discovering Posting and Reading Topics in Social Media*", W Gong, EP Lim, F Zhu
- [ApWeb'16] "*Detecting Community Pacemakers of Burst Topic in Twitter*", G Dong, W Yang, F Zhu, W Wang
- [WAIM'16] "*DPBT: A System for Detecting Pacemakers in Burst Topics*", G DONG, W YANG, F ZHU, W WANG
- [TKDE'16b] "*Efficient Online Summarization of Large-Scale Dynamic Networks*", Q Qu, S Liu, F Zhu, CS Jensen
- [TKDE'16a] "*TopicSketch: Real-time Bursty Topic Detection from Twitter*", W Xie, F Zhu, J Jiang, EP Lim, K Wang