RESEARCH STATEMENT

Jiawei Zhang, jzhan9@uic.edu
University of Illinois at Chicago
http://www.cs.uic.edu/∼jzhang2

Whether the people we follow in Twitter can be recommended as our potential friends in Facebook? How is the box office that US movies can achieve in China? How do weather and nearby points-of-interest (POIs) affect the traffic routes planned for vehicles? About the same information entities, e.g., social media users, imported foreign movies and vehicle routes, a large amount of information can actually be collected from various sources. These sources are usually of different varieties, like Facebook and Twitter, US and China movie markets, local weather and nearby POIs. Each information source provides a specific signature of the same entity from a unique underlying aspect. Effective fusion of these different information sources provides an opportunity for researchers and practitioners to understand the entities more comprehensively.

My research interest focuses on fusing multiple **large-scale** information sources of **diverse varieties** together [14, 15, 2, 10], and carrying out synergistic data mining tasks across these fused sources in one unified analytic [5, 6, 19, 1, 13, 11, 22, 12, 21, 3, 4]. Fusing and mining multiple information sources of large **Volumes** and diverse **Varieties** is a fundamental problem in **Big Data** studies. In my current research, I investigate the principles, methodologies and algorithms for synergistic knowledge discovery across multiple aligned information sources, and evaluate the corresponding benefits. I need to address challenges in the effective transfer of relevant knowledge across different aligned information sources, which depends upon not only the relatedness of these information sources, but also the target application problems, e.g. link prediction vs community detection vs information diffusion. I aim at developing general methodologies, which will be shown to work for a diverse set of applications, while the specific parameter settings can be learned for each application from the training data.

Current Research

My current research works on information fusion are mainly based on online social media data. Nowadays, to enjoy more social network services, people are usually involved in multiple online social networks simultaneously, such as Facebook, Twitter and Foursquare [19, 2]. Individuals usually have multiple separate accounts in different social networks, and discovering the correspondence between accounts of the same user (i.e., network alignment or user anchoring) [14, 15, 2, 10, 12, 4] will be an interesting problem. What's more, network alignment is also the crucial prerequisite step for many interesting inter-network synergistic knowledge discovery applications, like (1) inter-network recommendation [19, 5, 6, 12, 4], (2) mutual community detection [11, 1, 13, 3], and (3) cross-platform information diffusion [22, 21], with information from multiple networks simultaneously. These application tasks are fundamental problems in social network studies, which together with the network alignment problem will form the backbone of the multiple social network fusion learning ecosystem.

A. Heterogeneous Information Network Alignment

In online social networks, the social information generated by users' online social activities effectively indicates users' personal characteristics. Users' social information is usually of heterogeneous categories, involving both network structure information (like friendship and group membership) and attribute information (like user profile, location checkins and posts published online). Formally, we represent such a kind of networks as **heterogeneous information networks** [19]. Across social networks, the correspondence relationships between the common users' accounts are defined as **anchor links** [2]. In network alignment problems, we aim at inferring the potential anchor links between different social networks. I have done significant works on defining, formulating and solving the social network alignment problem based on different learning settings [14, 15, 2, 10, 12, 4].

A.1 Supervised Network Full and Partial Alignment: By treating the known anchor links as the training set, the network alignment problem can be formulated as a constrained supervised anchor link prediction problem. As shown in Figure 1, with the known anchor links and information in the heterogeneous social networks, we aim at inferring other potential anchor links (subject to *one-to-one* constraint) among users across networks. **To solve the problem, I propose a novel two-phase approach MNA: anchor link prediction + network matching [2]**. In inferring potential anchor links, a classifier is built with a set of features extracted for anchor links from the heterogeneous information across different networks. Based on the preliminary link inference results, the MNA approach further identifies the final alignment results by applying constrained stable matching to prune the redundant anchor links. MNA [2] performs very well for fully aligned social networks, in which all the users are anchor users. **To resolve the alignment problem for partially aligned networks, we propose another novel supervised network alignment algorithm named PNA in [10]**. PNA also involves two steps: anchor link inference + generic stable matching. Different from MNA, in the first step of PNA, conventional intra-network meta path concept is extended to the inter-network scenario based on the *anchor meta path* concept. A set of explicit and implicit features are extracted by PNA for anchor links based on the inter-network meta paths. Another significant difference from MNA is that, to handle the users who are not connected by anchor links, PNA allows users to stay isolated in the generic stable matching result across networks.

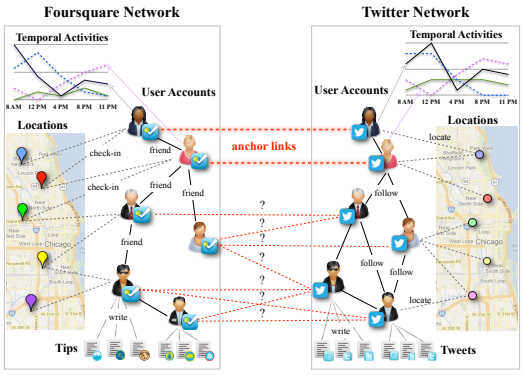


Figure 1: An example of social network alignment based on heterogeneous information.

A.2 Semi-supervised and Unsupervised Network Alignment: Anchor link training data is very expensive to obtain, since manual identification of common users' accounts across social networks is very difficult. To solve such a challenge, I propose to study the network alignment problem based on the *semi-supervised* and *unsupervised* learning setting respectively. **I have introduced a PU (Positive and Unlabeled) learning based network alignment framework, named CLF [12], to utilize the unlabeled anchor links in the model building**. In CLF, several new concepts, like *existence probability*, *formation probability* and *bridging probability*, about anchor links are introduced. By inferring anchor links' *bridging probability* from the training and validation sets, we can build models with a small number of partially observed anchor links together with the unlabeled anchor links to infer the network alignment results. **Furthermore, in the case that no known anchor links can be identified, I develop two other novel unsupervised frameworks for aligning multiple heterogeneous**

social networks, which are called M-NASA[14] and PCT [15] respectively. By studying the social network data, we observe that the shared users tend to have very similar social structure and attribute information in different social networks (especially for those of similar categories). By minimizing the structure loss and maximizing the attribute closeness of the inferred mappings, M-NASA and PCT formulate the network alignment problem as a constrained optimization problem.

A.3 Academic Impact Summary:

- MNA [2] focuses on supervised network full alignment problem, and PNA [10] studies supervised network partial alignment problem. MNA [2] is the first paper proposing the concepts of *multiple aligned social networks*, *anchor user* and *anchor link*. It is also one of the mostly cited paper in the proceedings of CIKM' 13, and has received 55 citations already.
- To solve the lack of training data problem, CLF [12] introduces the PU network alignment approach, and it has received 18 citations. PCT [15] focuses on social network co-alignment via different categories of shared information entities based on unsupervised learning setting. M-NASA [14] is the first work on aligning N networks simultaneously, which has been cited by 11 papers already in less one year. An extension of M-NASA [14] has been accepted by the AISC journal [20].

B. Recommendations across Aligned Networks

Generally, with the social activity data across multiple aligned social networks, we can acquire more comprehensive knowledge about users and their preferences. Based on such abundant prior knowledge, we can provide much better recommendation services for them. I have carried out a long-term social recommendation studies in the past years, and my works mainly focus on recommending offline POIs and online friends with information from multiple information sources. Formally, the friend and POI recommendation problems are formulated as the friendship and location link (i.e., the red dashed links to be inferred in Figure 2) prediction problems in our previous works [5, 6, 19, 12, 4].

B.1 POI Recommendation across Aligned Networks: A common problem encountered in recommendation services is the *information sparsity/cold start* problem when providing services for users with limited prior knowledge about their preferences. **To resolve such a severe problem, I propose to study the POI recommendation across multiple aligned social networks [6]**. For users with limited prior knowledge in one social network, we can transfer the abundant information about him/her from other aligned networks they are also involved in. Different from traditional transfer learning problems, the information is precisely transferred based on the identified *anchor links*. With the heterogeneous information available within and across the networks, a set of features indicating the users' personal interests in the POIs are extracted. Extensive empirical evaluations of the proposed model on real-world aligned network datasets show that our model can perform 36% better than traditional single-network based recommendation models measured by the AUC score.

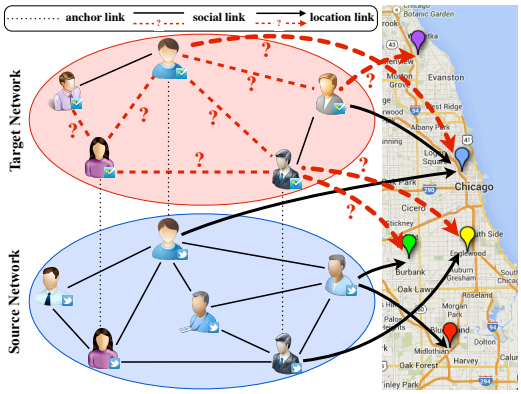


Figure 2: An example of recommendation across multiple aligned online social networks.

B.2 Friend Recommendation across Aligned Networks: Besides POIs, friend recommendation is another important service provided in online social networks. **Instead of targeting at regular users or networks, I study the friend recommendation problem for new users [5] and new networks [6, 19, 12] specifically**. For new users, we have limited knowledge about their preferences in terms of online friends [5]. Meanwhile, for new networks, the network itself is suffering from both the lack of training data and the shortage of information about the users [6, 19]. Thanks to the network alignment step, we can transfer both useful knowledge about the users as well as training data from other developed social networks to the target network via anchor links. To make full use of the heterogeneous information available across the aligned networks, a set of diverse features are extracted for friendship links in the model building. What's more, we observe that, for the user pairs which are not connected right now, some of them will be connected in the future but some will not. **Based on such an observation, we model the link prediction problem across aligned networks as a PU learning problem in [19, 12]** (in which the existing and non-existing links are treated as the positive and unlabeled instances). As to the domain difference problem, a novel network sampling based recommendation method is introduced in [5], and an automatic feature selection/weighting learning method is proposed in [19]. In our recent work [4], we propose to unify the link prediction problems with different cardinality constraints into a general framework, which can handle the links with *one-to-one*, *one-to-many* and *many-to-many* constraints at the same time.

B.3 Academic Impact Summary:

- TRAIL [6] works well on POI recommendation across aligned social networks. It addresses the cold start POI recommendation problem based on a set of features extracted for user-POI pairs across aligned social networks.
- SCAN-PS [5] and TRAIL [6] are the initial works studying cold start friend recommendation problems across aligned social networks. By now, they have received 28 and 42 citations already. MLI [19] introduces the novel inter-network meta path concept, and it has been cited by 40 research works. MLI [19] and CLF [12] propose the PU social link prediction problem. ITERCLIPS [4] unifies the prediction problems of links with different cardinality constraints.

C. Synergistic Community Detection across Aligned Networks

Information available across multiple aligned social networks provides more complete signals revealing the social community structures formed by people in the real world. Detecting the social community structures formed by users is also one of the key problem in my current research works. Based on multiple aligned social networks, we can transfer information from other mature networks to detect the community structure of emerging networks [11], as well as carrying out synergistic community detection tasks of multiple aligned social networks simultaneously [1, 13, 3].

C.1 Emerging Network Community Detection: For emerging networks with limited information, the problem of inferring the social community structure from the network is called the emerging network community detection problem [11]. Especially, when the network is brand new with no information about the users at all, the community detection will be named as the cold start community detection problem [11]. **To solve such problems, I develop a novel inter-network information transferring method, CAD, to propagate information from external source networks to the target network via diverse cross-network diffusion channels [11]**. In addition, to overcome the high time and space cost problems, an approximated scalable version of CAD is also introduced in [11], parameters involved in which can be effectively adjusted automatically by optimizing certain objective clustering metrics, e.g., mutual information or entropy.

C.2 Synergistic Community Detection across Aligned Networks: For networks with enough information to detect local community structures, information exchange among the aligned networks is also beneficial for the community detection problem as well. **I have also studied the synergistic mutual community detection problem across aligned networks to identify the community structure in each network simultaneously [13, 1, 3]**. The goal is to distill relevant information from other social networks to complement the knowledge directly derivable from each network to improve the detected communities, while preserving the distinct characteristics of each individual network at the same time. By minimizing the loss and discrepancy of the community detection results, the proposed framework, like MCD [13], SPMN [1] and MMC [3], can preserve the community structure characteristics and utilize external information (from other aligned networks) to refine and disambiguate the community structures of the common users [13, 1, 3]. Furthermore, a distributed version of the synergistic community detection framework SPMN is implemented on the Hadoop distributed computing cluster [1].

C.3 Academic Impact Summary:

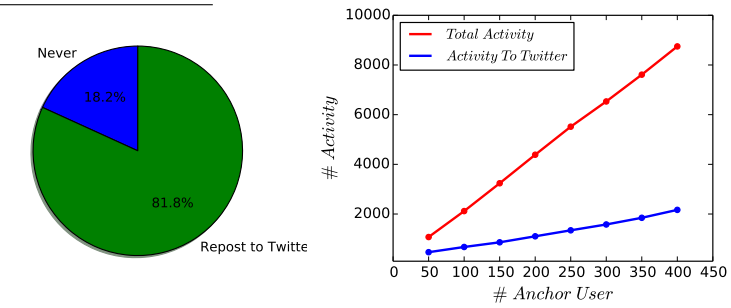
- CAD [11] introduces the emerging network clustering and cold start clustering problems. An improved version of CAD [11] has been submitted to the WWW journal.
- MCD [13], SPMN [1] and MMC [3] are the initial research works on large-scale concurrent clustering of multiple aligned social networks. SPMN [1] has been downloaded more than 80 times and cited by 18 recent research papers.

D. Information Diffusion across Aligned Networks

In addition, the formulation of multiple aligned heterogeneous social network provides researchers the opportunity to study the information diffusion process across different social sites.

D.1 Information Diffusion across Aligned Social Networks:

Via the shared users, information can be effectively propagated across social networks via social activities like tweet/tip repost. With a thorough analysis about the Twitter and Foursquare social network datasets, inter-network information repost is a common practice in the real world, and 81.8% of the shared users have ever repost their tips and checkins from Foursquare to Twitter (see Figure 3(a)). In addition, the reposted contents actually account for about 25% of the users' total activities in Foursquare (see Figure 3(b)). **To model such an important observation, we propose a Multi-source Multi-channel information diffusion model to depict how information propagates across aligned networks in [22]**. Based on the diffusion model, various interesting information diffusion related problems can be studied across aligned networks, like cross-network viral marketing [22], and inter-network tipping user discovery [21].



(a) Foursquare users' reposting (b) cross-network reposted activities

Figure 3: Cross-network information propagation analysis.

D.2 Academic Impact Summary:

- MMN [22] introduces the information diffusion problem across aligned social networks, which has been downloaded 1,585 times and cited by 16 papers. CONFORM [21] defines the tipping point concept and the tipping user discovery problem.

Future Agenda

As more data is being generated in different disciplines, the needs for effective and efficient fusion and synergistic knowledge discovery algorithms will keep increasing steadily.

A. Long-term Goal: Theoretical Fusion Learning and Mining

Academic Goal: My future research interests are directed towards a long-term goal of making sense of big data in a holistic perspective. In the future, I plan to propose the concept of *fusion learning* and focus on fusing data from different sources for integral knowledge discoveries. To achieve performance guarantees, a complete theoretic framework for fusion learning and mining studies is to be introduced. Theoretic performance bounds will be derived for input data following different distributions. Robust information source selection and weighting methods will also be studied to resolve the domain difference problem. Based on the theoretic foundation, different fusion learning frameworks will be proposed to deal with various categories of datasets, ranging from traditional feature vector representation data, to graph/network data, text data, time series data and image/video data. Besides aligning social networks, more different information fusion algorithms will be proposed to unify different types of input data in the future. In addition, more diverse application problems will be studied based on the unified fused data, which will lead to more opportunities and developments.

Industrial Goal: Besides the academic goals, I also plan to introduce the concepts and algorithms of fusion learning to the industry to apply it in joining both (1) different departments in big companies, and (2) big companies with small startups. On the premise that privacy concerns are not violated, by connecting various company internal departmental information together, such as Google Search, Gmail, Google Maps and Youtube, big companies like Google can greatly improve the services provided by these departments in the company. Furthermore, forming alliance between big companies and emerging startups can lead to a win-win payoff for both of them via information sharing. With the abundant information shared from big companies, startups can survive and become prosperous much easier. Meanwhile, big companies can also utilize their data advantages to tap into novel or disruptive technologies to keep their leadership positions in the future industrial evolutions.

B. Short-term Goal: Scalable Fusion and Mining Applications of Various Multi-Source Data

Scalable Fusion Learning Algorithms: Data generated nowadays is usually of very large scale, and fusion of such big data from multiple sources together will render the problem more challenging. For instance, the online social networks (like Facebook) usually involve millions even billions of active users, and the social activity data generated by these users in each day will consume more than 600 TB storage space (in Facebook). One of the major short-term objective of my research is to develop scalable data fusion and mining algorithms that can handle such a **large volume** (of **big data**) challenge. One tentative approach is to develop information fusion algorithms based on distributed platforms, like Spark and Hadoop [1], and

handle the data with a large distributed computing cluster. Another method to resolve the scalability challenge is from the model optimization perspective. Optimizing existing learning models and proposing new approximated learning algorithms with lower time complexity are desirable in my future research projects. In addition, I'm also interested in applying deep learning models to fuse and mine the large-scale datasets.

Multiple Sources Fusion and Mining: Current research works on multiple source data fusion and mining mainly focus on aligning entities in one single pair of data sources (i.e., two sources), where information exchange between the sources mainly rely on the anchor links between these aligned entities. Meanwhile, when it comes to the fusion and mining of multiple (more than two) sources, the problem setting will be quite different and become more challenging. For example, in the alignment of multiple (more than two) networks, the transitivity property of the inferred anchor links needs to be preserved [14]. Meanwhile, in the information transfer from multiple external aligned sources to the target source, the information sources should be weighted differently according to their importance. Therefore, the **diverse variety** of the multiple sources will lead to more research challenges and opportunities, which is also a great problem in **big data** studies. I will introduce new information fusion and mining algorithms for the multi-source scenerios.

Broader Fusion Learning Applications: Besides the research works on social network datasets, I'm also interested in the application of fusion learning and mining algorithms on other categories of datasets, like enterprise internal data [16, 7, 18, 17], geo-spatial data [8, 21, 9], knowledge base data, and pure text data. I have some prior research works on fusing enterprise context information sources, like enterprise social networks, organizational chart and employee profile information. During my PhD work, I have studied several interesting problems, like organizational chart inference [16], enterprise friendship link prediction [7], information diffusion at workplace [18] and enterprise employee training [17], based on the fused enterprise internal information. In the future, I'm interested in applying the fused enterprise internal information sources in other application problems, such as expert location and project team formation. For the offline spatial data, I'm interested in analyzing the correlation of different traveling modalities (like shared bicycles [8, 21, 9], bus and metro train) with the city zonings. I'm also interested in fusing multiple knowledge bases, like Douban and IMDB, for knowledge discovery and truth finding.

C. Fundings and Collaborations

During my PhD studies, I have actively participated in the research grant proposal writing. Currently, under the supervision of Prof. Philip S. Yu (the PI), I have contributed significantly to a successful NSF information fusion and mining proposal (NSF III-1526499), which lead to a 3-year $500,000 funding to UIC and supported my PhD studies. As a major contributor to the proposal, I'm responsible for defining new information fusion problems and designing new algorithms to solve the problems. In the future, to continue my research in this area, I plan to write several grant proposals to apply for more research fundings from NSF and NIH. In addition, in the past years, close collaboration with researchers in both academia and industry have shaped my research style a lot and brought about lots of novel ideas. With Dr. Yuanhua Lv (from Microsoft Research), we have studied the problems to fuse and mine enterprise internal information sources. With Dr. Charu C. Aggarwal (from IBM T. J. Watson Research), we have studied the signed network mining problems by fusing links belonging to different polarities into an integrated mining framework. With Dr. Jianhui Chen (from Yahoo! Research), we have focused on optimizing fusion learning models to lower down their time complexities. In the near future, I will continue the close collaboration with more researchers both within and outside my area in carrying out research projects about *data fusion and knowledge discovery*.

Selected References

[1] Songchang Jin, **Jiawei Zhang**, Philip S. Yu, Shuqiang Yang, and Aiping Li. Synergistic partitioning in multiple large scale social networks. In *Proceedings of the 2014 IEEE International Conference on Big Data*, BigData, 2014.

[2] Xiangnan Kong, **Jiawei Zhang**, and Philip S. Yu. Inferring anchor links across multiple heterogeneous social networks. In *Proceedings of the 22nd ACM International Conference on Conference on Information and Knowledge Management*, CIKM, 2013.

[3] Weixiang Shao, **Jiawei Zhang**, Lifang He, and Philip S. Yu. Multi-source multi-view clustering via discrepancy penalty. In *Proceedings of the International Joint Conference on Neural Networks*, IJCNN, 2016.

[4] **Jiawei Zhang**, Jianhui Chen, Junxing Zhu, Yi Chang, and Philip S. Yu. Link prediction with cardinality constraints. In *Proceedings of the 10th ACM International WSDM Conference on Web Search and Data Mining*, WSDM, 2017.

[5] **Jiawei Zhang**, Xiangnan Kong, and Philip S. Yu. Predicting social links for new users across aligned heterogeneous social networks. In *IEEE 13th International Conference on Data Mining*, ICDM, 2013.

[6] **Jiawei Zhang**, Xiangnan Kong, and Philip S. Yu. Transferring heterogeneous links across location-based social networks. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM, 2014.

[7] **Jiawei Zhang**, Yuanhua Lv, and Philip S. Yu. Enterprise social link prediction. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*, CIKM, 2015.

[8] **Jiawei Zhang**, Xiao Pan, Moyin Li, and Philip S. Yu. Bicycle-sharing system analysis and trip prediction. In *Proceedings of the 17th International Conference on Mobile Data Management*, MDM, 2016.

[9] **Jiawei Zhang**, Xiao Pan, Moyin Li, and Philip S. Yu. Bicycle-sharing systems expansion: Station re-deployment through crowd planning. In *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information System*, SIGSPATIAL, 2016.

[10] **Jiawei Zhang**, Weixiang Shao, Senzhang Wang, Xiangnan Kong, and Philip S. Yu. Pna: Partial network alignment with generic stable matching. In *Proceedings of the 16th IEEE International Conference on Information Reuse and Integration*, IRI, 2015.

[11] **Jiawei Zhang** and Philip S. Yu. Community detection for emerging networks. In *Proceedings of the 15th SIAM International Conference on Data Mining*, SDM, 2015.

[12] **Jiawei Zhang** and Philip S. Yu. Integrated anchor and social link predictions across partially aligned social networks. In *Proceedings of the International Joint Conference on Artificial Intelligence*, IJCAI, 2015.

[13] **Jiawei Zhang** and Philip S. Yu. Mcd: Mutual clustering across multiple social networks. In *Proceedings of IEEE BigData Congress*, BigData Congress, 2015.

[14] **Jiawei Zhang** and Philip S. Yu. Multiple anonymized social networks alignment. In *Proceedings of the 15th IEEE International Conference on Data Mining*, ICDM, 2015.

[15] **Jiawei Zhang** and Philip S. Yu. Pct: Partial co-alignment of social networks. In *Proceedings of the 25th World Wide Web Conference*, WWW, 2016.

[16] **Jiawei Zhang**, Philip S. Yu, and Yuanhua Lv. Organizational chart inference. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD, 2015.

[17] **Jiawei Zhang**, Philip S. Yu, and Yuanhua Lv. Enterprise employee training via project team formation. In *Proceedings of the 10th ACM International WSDM Conference on Web Search and Data Mining*, WSDM, 2017.

[18] **Jiawei Zhang**, Philip S. Yu, Yuanhua Lv, and Qianyi Zhan. Information diffusion at workplace. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management*, CIKM, 2016.

[19] **Jiawei Zhang**, Philip S. Yu, and Zhi-Hua Zhou. Meta-path based multi-network collective link prediction. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD, 2014.

[20] **Jiawei Zhang**, Qianyi Zhan, and Philip S. Yu. Concurrent alignment of multiple anonymized social networks with generic stable matching. In Thouraya Bouabana-Tebibel and H Stuart Rubin, editors, *Theoretical Information Reuse and Integration*. Springer International Publishing, 2016.

[21] Qianyi Zhan, **Jiawei Zhang**, Xiao Pan, Moyin Li, and Philip S. Yu. Discover tipping users for cross network influencing. In *Proceedings of IEEE 17th International Conference on Information Reuse and Integration*, IRI, 2016.

[22] Qianyi Zhan, **Jiawei Zhang**, Senzhang Wang, Philip S. Yu, and Junyuan Xie. Influence maximization across partially aligned heterogenous social networks. In *Pacific Asia Knowledge Discovery and Data Mining*, PAKDD, 2015.