# RESEARCH STATEMENT

Chuxu Zhang, czhang11@nd.edu
University of Notre Dame
https://chuxuzhang.github.io/

## Overview

In today's information and computational society (see Fig. 1), complex systems (e.g., e-commerce system, online social network, enterprise network, transportation network, chemical synthesis, biomedicine) are often associated with heterogeneous (multi-modal) data (e.g., structural relation, unstructured text/image, temporal context). The heterogeneous data provide opportunities for researchers and practitioners to understand complex systems more comprehensively but also pose challenges to discover knowledge from them. The challenges come from not only the complexity and heterogeneity of the data for investigation but also the requirement of the target problem to be solved. Besides the difficulty of extracting useful information from the complex data, it is hard to fuse the extracted knowledge in a unified manner so as to facilitate various underlying applications. Can we develop artificial intelligence solutions to extract, represent, fuse knowledge from heterogeneous data so as to tackle various challenges in complex systems?

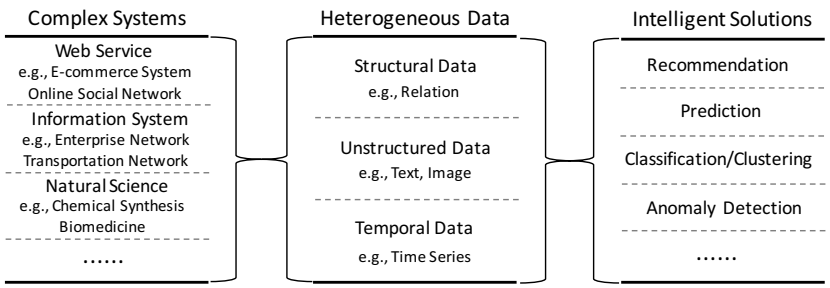


Figure 1: Overview of my research: turning heterogeneous data into useful knowledge, upon what developing artificial intelligence solutions for a diverse set of applications in complex systems across different disciplines.

**The goal of my research is to harness the power of heterogeneous data, turn them into useful knowledge, develop artificial intelligence solutions based on the extracted knowledge for a diverse set of real-world applications in complex systems across different disciplines**. Successful AI solutions should be able to solve two fundamental research questions brought forth by the above challenges: (1) How to extract and represent useful information from the data of heterogeneous structure (e.g., multi-typed objects interconnected by multi-typed relations), or multi-modal/source (e.g., structural data, unstructured data, temporal data), or both? (2) How to fuse the extracted knowledge in customized machine learning models for solving target problems (e.g., recommendation, prediction, classification/clustering, anomaly detection) across different disciplines (e.g., web service, information system, natural science)? I have been developing a series of methodologies and algorithms to answer these questions, which have been deployed and validated in several research topics (e.g., personalization/recommendation, representation learning). **My research has led to over 10 papers in the top conferences of artificial intelligence (e.g., AAAI, IJCAI), data science (e.g., KDD), and web/information systems (e.g., WWW). As of 11/2019, my google scholar citation is over 400**. In the following sections, I will highlight my current research and present my future agenda.

## Current Research: Learning From Heterogeneous Networks

During my Ph.D. study, I am fortunate to join Network Science Collaborative Technology Alliance[1] and be supported by U.S. Army Research Lab (ARL). My dissertation research - learning from heterogeneous networks, is an important part of ARL Network Sciences Project[2] and focuses on developing AI solutions for solving different problems in heterogeneous networks. Nowadays, people are deeply involved in various online web services such as social media (e.g., Facebook), online shopping (e.g., Amazon), academic search (e.g., Google Scholar), etc. Those complex systems are usually modeled

[1] http://www.ns-cta.org/ns-cta-blog/
[2] https://www.arl.army.mil/www/default.cfm?page=391

as heterogeneous networks associated with heterogeneous data, representing multi-typed nodes interconnected by multi-typed edges as well as multi-modal/source contents in nodes and edges. For example, the online shopping data is of heterogeneous structure with <shop, item, user, brand, etc.> nodes and <item in shop, user buys item, item of brand, etc.> edges/relations. In addition, nodes (e.g., items) are associated with heterogeneous content (e.g., text description, picture, timestamp). It is hard for individuals to seek personal needs in those systems, especially when the candidate set is large (e.g., millions of items). Thus developing intelligent solutions to automatically filter information for individuals and provide personalized services to them is important. In addition, the abundant heterogeneous information in those systems requires both a domain understanding and large exploratory search space when doing feature engineering activities of customized machine learning models for different purposes. Therefore, developing intelligent solutions to generalize the feature engineering activity through automating the discovery of feature representation for various tasks is crucial. Accordingly, my current research tackles problems from two perspectives: (a) personalization in heterogeneous networks; and (b) heterogeneous network representation learning. Most of them are covered in our recent tutorial [3].

## A. Personalization in Heterogeneous Networks

The problem of personalization aims at automatically recommending suitable objects (e.g., items) to target entities (e.g., users) in the system. As many web services (e.g., e-commerce systems) are modeled as heterogeneous networks, extracting and fusing useful information from such complex structure would benefit recommender system designs in those services. Driving by these facts, we have developed a series of works in this direction. To address the data sparsity and cold-start issue, we proposed **CUNE** [8], a collaborative user identification model which leverages user-item bipartite structure to identify top similar users of each user. The identified users are then incorporated into standard recommendation models (e.g., matrix factorization, pairwise ranking) for improving performances in both rating prediction and item ranking tasks. We further proposed **WalkRanker** [12] for elevating item ranking quality by incorporating multiple user-item relations into a unified pairwise ranking model. Moreover, to address sequence modeling challenge (i.e., recommending next item), we developed **MARank** [11] to improve sequential recommendation performance by unifying both individual- and union-level item sequential correlations in preference ranking model.

Besides of investigating about recommender systems, we have also conducted some works for personalization in the online academic service systems, which tackle the data of heterogeneous academic networks with text content in the nodes. We proposed **Camel** [2] for author identification (i.e., predicting authors of anonymous papers), which is useful for reviewer recommendation. The model (Fig. 2) is briefly illustrated as follows.

- **Information extraction and representation.** Camel uses deep content encoding (i.e., $f(p)$) and latent feature (i.e., $q$) to represent paper and author. Besides, Camel extracts direct author-paper relations and employs meta-path walks to capture multiple indirect author-paper correlations in the network.
- **Knowledge fusion.** Camel leverages distance ranking metric to model direct author-paper relations (i.e., pulling correlated authors inward while pushing uncorrelated ones outward to the target paper in gradient direction), and further augments the framework through a regularization term of indirect author-paper correlations extracted from the heterogeneous structure.

By jointly optimizing distance ranking metric and heterogeneous structure regularization, Camel is able to effectively predict correlated/true authors of anonymous papers. Besides Camel, we also developed **TSR** [9], a task-guided and semantic-aware ranking model, to effectively recommend academic papers to researchers, especially for whom with little background knowledge of the field.

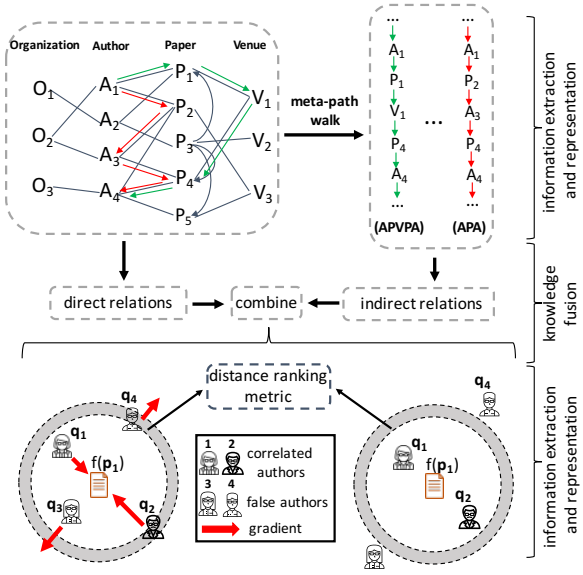


Figure 2: Illustration of Camel [2]: knowledge extraction and fusion for personalization in heterogeneous networks.

### B. Heterogeneous Network Representation Learning

The purpose of heterogeneous network representation learning is to automate the discovery of meaningful vector representation for each node in the network so as to reduce labor-consuming feature engineering activity and facilitate various downstream tasks. We have pursued this direction and have generated several research outputs in this area. We proposed **SHNE** [6], i.e., semantic-aware heterogeneous network embedding, to learn node embeddings in heterogeneous networks with text information (e.g., academic network with text content in nodes). SHNE addresses the challenge of extracting and fusing both structural closeness and semantic correlations by integrating node content (i.e., text) as a deep encoding function into the heterogeneous structure embedding framework, and elevates performances in various network mining tasks.

Following the success of SHNE and in order to achieve a bigger scenario goal that learning node representations in heterogeneous networks with multi-modal/source content in nodes (e.g., online shopping/reviewing networks with title/description text, review text, image/picture, attribute information in the nodes), we further developed **HetGNN** [4], a heterogeneous graph neural network framework. The design of HetGNN (Fig. 3) is briefly illustrated as follows.

- **Information extraction and representation.** HetGNN extracts different types of neighboring nodes of the target node (i.e., node $a$) by a random walk based approach over heterogeneous structure and selects the most frequently visited nodes as correlated nodes of the target node. In addition, HetGNN employs cross-domain deep learning techniques to obtain feature representations of multi-modal content in each node (e.g., natural language model for encoding text feature, computer vision model for encoding image feature).
- **Knowledge fusion.** HetGNN aggregates multi-modal content features of each node and further aggregates representations of heterogeneous correlated node types for the target node by different deep learning modules. Therefore, HetGNN fuses the knowledge extracted from all correlated nodes and obtains the final representation of the target node.

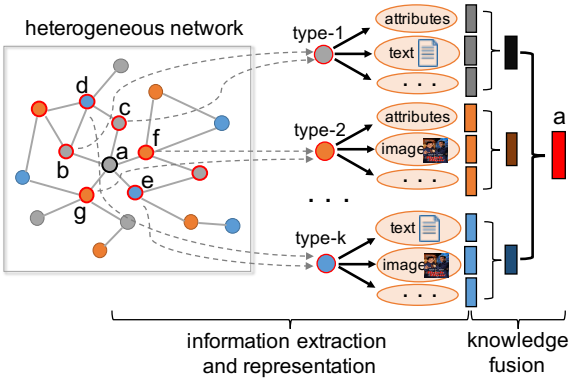


Figure 3: Illustration of HetGNN [4]: knowledge extraction and fusion for representation learning in heterogeneous networks.

HetGNN has been deployed in a number of web service system data and performs very well for various downstream tasks including link prediction, node recommendation, node classification/clustering, and embedding visualization. In addition to SHNE and HetGNN, we also proposed **FSRL** [7], a relational representation learning model in knowledge graphs (a case of heterogeneous network with abundant node/relation types). FSRL considers heterogeneity of both neighboring entities and relations in learning target entity's embedding, which elevates link prediction performance.

## Research Vision: Future Agenda

My future research agenda is directed towards a long-term goal of developing artificial intelligence algorithms and softwares, addressing the challenges of heterogeneity and multiple modality in data, that carry an interdisciplinary impact. My goal will be to develop effective, efficient, and interpretable solutions.

1. **In-depth analysis and broader applications.** My current research has explored the heterogeneous data associated with heterogeneous structure or multi-modal content or both. Most of complex systems are also associated with temporal data (e.g., time series, spatial-temporal context) which is useful for investigating system dynamics and improving the predictive analysis. I have done some work for time series and spatial-temporal data analysis [5, 1], and can not wait to leverage those information to further improve the state-of-the-art for different problems in heterogeneous data. In addition, I currently work for the NSF Center for Computer Aided Synthesis[3] which aims to

[3] https://ccas.nd.edu/

use quantitative, data-driven approaches to make synthetic chemistry more predictable. Many of chemical synthesis problems can be transformed to machine learning tasks in heterogeneous data. For example, predicting the product of a chemical reaction is to infer edge/relation change (i.e., bond change) in heterogeneous networks (i.e., molecular graphs with multi-typed attributed atoms and bonds). I am excited to use my knowledge to develop intelligent solutions for various problems across different disciplines (e.g., natural science).

2. **Learning with small labeled data.** Missing or lacking ground-truth labels is common in heterogeneous data and it is often expensive to collect such labels. For example, e-commerce systems and knowledge graphs face cold-start issue. The product collection of chemical reactivity spends a lot of money for human-laboring and experimental resource. It is practically significant to create efficient machine learning models to solve the challenges of small labeled data. My prior works FSRL [7] and GFL [10] tackle limited supervisory labels (i.e., few-shot relational entity pairs) in graph data. Depart from these study, I will investigate more techniques (e.g., meta-learning) for developing efficient solutions using small labeled data.
3. **Interpretable learning.** The current machine learning algorithms for heterogeneous data are mostly lack of explainability. For example, the link prediction methods predict potential connection between two nodes yet do not show why such connection is inferred. The recommendation models suggest items to users while do not tell users why such items are recommended. Equipping the machine learning algorithms with explainable capability will indeed bring benefits in providing a transparent, trustworthy, and effective solutions. I am interested in leveraging advanced techniques (e.g., reinforce learning) to develop interpretable solutions.

**Collaboration and Funding.** I am supported by NSF and ARL. In addition, previously I worked at Microsoft Research and NEC Labs America as research intern. I have close collaboration with a good number of researchers in both academia and industry. Moreover, I actively help my advisor in grant proposal writings during Ph.D. study. In the future, I will actively write proposals to apply for research grants from multiple funding agencies (e.g., NSF, ARL, DARPA, NIH) and industry companies.

# References

[1] Chao Huang, **Chuxu Zhang**, Jiashu Zhao, Xian Wu, Nitesh V Chawla, and Dawei Yin. Mist: A multiview and multimodal spatial-temporal learning framework for citywide abnormal event forecasting. In *Web Conference (World Wide Web Conference), WWW*, 2019.

[2] **Chuxu Zhang**, Chao Huang, Lu Yu, Xiangliang Zhang, and Nitesh V Chawla. Camel: Content-aware and meta-path augmented metric learning for author identification. In *Web Conference (World Wide Web Conference), WWW*, 2018.

[3] **Chuxu Zhang**, Meng Jiang, Xiangliang Zhang, and Nitesh V Chawla. Multi-modal network representation learning: methods and applications. In *SIAM International Conference on Data Mining, SDM (tutorial)*, 2020.

[4] **Chuxu Zhang**, Dongjin Song, Chao Huang, Ananthram Swami, and Nitesh V Chawla. Heterogeneous graph neural network. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD*, 2019.

[5] **Chuxu Zhang**, Dongjin Song, Chen Yuncong, Xinyang Feng, Cristian Lumezanu, Wei Cheng, Jingchao Ni, Bo Zong, Haifeng Chen, and Nitesh V Chawla. A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data. In *AAAI Conference on Artificial Intelligence, AAAI*, 2019.

[6] **Chuxu Zhang**, Ananthram Swami, and Nitesh V Chawla. Shne: Representation learning for semantic-associated heterogeneous networks. In *ACM International Conference on Web Search and Data Mining, WSDM*, 2019.

[7] **Chuxu Zhang**, Huaxiu Yao, Chao Huang, Meng Jiang, Zhenhui Li, and Nitesh V Chawla. Few-shot knowledge graph completion. In *AAAI Conference on Artificial Intelligence, AAAI*, 2020.

[8] **Chuxu Zhang**, Lu Yu, Yan Wang, Yan Wang, and Xiangliang Zhang. Collaborative user network embedding for social recommender systems. In *SIAM International Conference on Data Mining, SDM*, 2017.

[9] **Chuxu Zhang**, Lu Yu, Xiangliang Zhang, and Nitesh V Chawla. Task-guided and semantic-aware ranking for academic author-paper correlation inference. In *International Joint Conferences on Artificial Intelligence, IJCAI*, 2018.

[10] Huaxiu Yao, **Chuxu Zhang**, Ying Wei, Meng Jiang, Suhang Wang, Junzhou Huang, Nitesh V Chawla, and Zhenhui Li. Graph few-shot learning via knowledge transfer. In *AAAI Conference on Artificial Intelligence, AAAI*, 2020.

[11] Lu Yu, **Chuxu Zhang**, Shangsong Liang, and Xiangliang Zhang. Multi-order attentive ranking model for sequential recommendation. In *AAAI Conference on Artificial Intelligence, AAAI*, 2019.

[12] Lu Yu, **Chuxu Zhang**, Shichao Pei, Guolei Sun, and Xiangliang Zhang. Walkranker: A unified pairwise ranking model with multiple relations for item recommendation. In *AAAI Conference on Artificial Intelligence, AAAI*, 2018.