Proposal Form
PhD Thesis / Research residency stay
China Scholarship Council (CSC)/ENS Rennes
Call for projects 2022

FIELD: Computer Science

THESIS / RESEARCH SUBJECT TITLE

Graph-based Data Mining for Social Network Analysis

Type of PhD: please tick the appropriate box:

- Joint PhD/cotutelle (leading to a double diploma) : YES ☐ or NO ☑
- Regular PhD (leading to a single French diploma) : YES ☑ or NO ☐
- Research residency visit (for students enrolled at a Chinese institution who will be invited to a French institution to carry out a mobility period) : YES ☐ or NO ☑

Name of French doctoral school (if applicable): University of Rennes 1

Name of host laboratory and research team: IRISA, DRUID team

Website: http://www.irisa.fr/druid

Name of French Supervisor: Arnaud Martin

Name of PhD director-s in Chinese university (if applicable):

- Name: Laurence T. Yang
- Position: Chair Professor
- E-mail: ltyang@gmail.com

If previous collaborations with the Chinese co-director/university, please elaborate:

We have an active research collaboration with Prof. Laurence T. Yang of Hainan University, China, co-director of this proposal. In partnership, we have done some data mining work based on the theory of belief functions (TBF) in modeling and managing uncertain and imprecise information, for example, caused by missing values in incomplete data. We also developed a flexible and fast framework to reduce the complexity in clustering based on the TBF. As a result, some have been published in top journals, such as ***IEEE Trans. Knowledge and Data Eng.***, ***Knowl.-Based Syst.***, ***Inf. Sci.***.

Moreover, we are currently supervising Zhe LIU (Ph.D. applicant of this proposal) for incomplete data mining and analysis in Cyber-Physical-Social Systems (CPSS), aiming to address the uncertainty and

imprecision in high-dimensional data based on the TBF. This work has been submitted to the top journal (***IEEE Trans. Fuzzy Syst.***). Besides, we are also studying fuzzy data mining in CPSS, especially in social systems.

Furthermore, as the general chair and steering committee chair, the two teams are jointly organizing the 7th International Conference on Data Science and Systems, which will be held during December 20-22, 2021, in Hainan, China.

In recent years, with the development of information technology, social networks have become one of the focuses. However, real-world network data usually has a large amount of uncertain information. Therefore, as a practical theory to deal with uncertain information, the TBF is of great significance in developing flexible frameworks for analyzing the uncertainty and imprecision in social networks.

Our motivation is to establish a long-term research collaboration with Hainan University in data science, especially mining social network data in CPSS.

Research proposal abstract (1500 words max.):

The last decades have seen a surge of interest in graph-based network data mining problems that consider different forms of uncertainties. Successful graph-based approaches mainly rely on classical theories, such as probability theory, to deal with the uncertain information in real-world graphs. However, in the recent few years, new uncertainty theories have emerged. In particular, the theory of belief functions (TBF) [1] has received growing interest due to its richer representation of uncertainty and imprecision compared to probability theory and its higher ability to combine pieces of information.

This Ph.D. work will focus on graph-based data mining for social network analysis. Online social networks allow only limited access to their data which generates more imbalance, imprecision, and uncertainty for the graph-based social network analysis fields [2,3].

The primary areas to investigate are user (symmetric or asymmetric) clustering based on uncertain heterogeneous information fusion. Specifically, in the first phase, we intend to design a graph-based clustering algorithm that applies to balanced and imbalanced social network data. Then, in the second phase, we plan to combine this first line of work with the possibility of grouping users in the network using multiple information (possibly heterogeneous and highly conflicting) from the complementary sources.

Over the longer term, the goal is to develop a versatile and flexible graph-based framework for addressing uncertainty in social network analysis.

Detailed thesis topic

1 Graph-based social networks asymmetric clustering based on uncertain information

With the pervasive use of social networks, people have become more reliant on social media to get information and news. It also offers interesting research opportunities for analyzing the behavior and

interactions between/among users. However, online social networks only allow limited access, which will bring imbalance and uncertainty to the field of graph-based social network analysis, and this has not attracted enough attention. Thus, how to meet the needs of "small" and "large" communities according to the opinions of their like-minded neighbours has become a problem worthy of study, which can also be seen as graph-based asymmetric clustering problems on social networks.

Most graph-based social data clustering techniques ignore the adverse effects of imbalanced data and focus only on symmetrical data. This project will focus on how to solve the imbalance of social data to establish a graph-based clustering algorithm that can improve the balance and imbalance of social data and improve the efficiency of the graph-based social data algorithm. The TBF is widely used for graph-based social networks clustering and has already been adopted to describe the uncertain community structure in graphs [4,5]. However, the current evidential clustering methods have high complexity. Thus, it is of great value if an efficient evidential clustering method can be designed for sizeable asymmetric data sets.

Another problem we will consider in graph-based asymmetric clustering is that there may be some potential prior information in social networks, such as pairwise constraints and some pre-labeled nodes. This knowledge can be obtained from websites, primarily online crowdsourcing platforms [6]. We will study how to use a small amount of supervised information and group data into "small" and "large" groups by means of multiple clustering and merging to improve the clustering accuracy as much as possible and design an online graph-based social networks detection approach that can interact with the crowdsourcing applications, which will enable us to have a deeper understanding of the network data set, and help us to make better use of relevant fields.

The main challenges in this topic are as follows:

(1) How to cluster "small" communities reasonably and exactly while clustering "large" communities into multiple categories?

(2) How to merge the "small" communities who belong to the same "large" communities and separated because of multi-clustering reasonably and efficiently?

2 Active social user clustering based on heterogeneous information fusion

Nowadays, interactions between users are not only limited to social relations, but also to some activities, such as reading and writing. Thus, multiple and complementary information sources are available for characterizing users [7]. One task that could benefit from the integration of those multiple sources is community detection, which can also be seen as clustering problems on graph data sets.

Most graph clustering techniques have disregarded the effect of information aggregation and continue to focus only on the topological structure of graphs. This project will focus on how to take advantage of the multiple social and content-based information originated in social networks for improving the performance of the community detection algorithms based on the first stage.

The content information included in the network can be attributes, user behaviors, etc, which could be crawled from the websites. Thus, it will bring about a large amount of uncertainty and imprecision. The information from different views might be of heterogeneous and high conflict. In this part, we will study the user clustering problems in social networks from information fusion using the TBF.

The main challenges in this topic are as follows:

(1) How to transform heterogeneous information and integrate them in homogeneous framework?

(2) How to establish an effective high conflict fusion method in social network information?

Reference

[1] G. Shafer, *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[2] E. Adar and C. Re, Managing Uncertainty in Social Networks. *IEEE Data Engineering Bulletin*, vol. 30, pp. 15-22, 2007.

[3] P. Chunaev, Community Detection in Node-attributed Social Networks: A Survey. *arXiv e-prints*, 2019, arXiv: 1912.09816.

[4] K. Zhou, A. Martin, Q. Pan and Z. G. Liu, Median Evidential C-means Algorithm and its Application to Community Detection. *Knowledge-Based Systems*, vol. 74, pp. 69–88, 2015.

[5] K. Zhou, A. Martin, Q. Pan and Z. G. Liu, SELP: Semi-supervised Evidential Label Propagation Algorithm for Graph Data Clustering. *International Journal of Approximate Reasoning*, vol. 92, pp. 139-154, 2018.

[6] Y. Yan, R. Rosales, G. Fung and J. G. Dy, Active Learning from Crowds. *International Conference on Machine Learning*, Bellevue, Washington, USA, 2011, pp. 1161–1168.

[7] G. Guo, J. Zhang and N. Yorke-Smith, Leveraging Multiviews of Trust and Similarity to Enhance Clustering-based Recommender Systems. *Knowledge-Based Systems*, vol. 74, pp. 14-27, 2015.

Publications of the laboratory in the field (max 5):

[1] Z. W. Zhang, Z. Liu, A. Martin, Z. G. Liu and K. Zhou, Dynamic Evidential Clustering Algorithm. *Knowledge-Based Systems*, vol. 213, 2021. doi: 10.1016/j.knosys.2020.106643.

[2] Z. W. Zhang, H. P. Tian, L. Z. Yan, A. Martin and K. Zhou, Learning a Credal Classifier with Optimized and Adaptive Multiestimation for Missing Data Imputation. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2021. doi: 10.1109/TSMC.2021.3090210.

[3] Y. R. Zhang, T. Bouadi, Y. W. Wang and A. Martin, A Distance for Evidential Preferences with Application to Group Decision Making. *Information Sciences*, vol. 568, pp. 113-132, 2021.

[4] Z. W. Zhang, A. Martin, Z. G. Liu, K. Zhou and Y. R. Zhang Fast Unfolding of Credal Partitions in Evidential Clustering. *International Conference on Belief Functions*, Springer, Cham, 2021, pp. 3-12.

[5] K. Zhou, M. Guo and A. Martin, Evidential Clustering based on Transfer Learning. *International Conference on Belief Functions*, Springer, Cham, 2021, pp. 56-65.

Date:

Signature of the French Supervisor: