# Mining Overlapping Clusters in Networks

Recent advances in social media and social networks have made clustering on very large networks a vibrant and important research topic. In many real world applications, we require clusters to be overlapping [22]. For example, the clusters in a social network are overlapping groups of friends [4], and clusters in scientific collaboration network are overlapping sets of authors sharing some common research interests [2],etc.. Despite a rich literature on overlapping clustering in networks, most the existing works suffer from the following shortcomings.

- **Weighted Networks.** Weighted networks are ones with links associated with weights representing the strength of the relationship among the nodes. In addition, in many context, the weight can be either positive or negative, depending on the relationship between the nodes. For example, in social networks, a positive link represents *friend* and *support* relationship, while a negative link represents *foe* or *against* relationships [11, 8]. In trust networks, positive and negative links reflect the trust and distrust relationship respectively [10]. Although link weight provides more insights about the network structure [3], there are only very few clustering methods that can mine overlapping clusters from weighted networks [6, 13]. Moreover, these methods do not consider negative links, hence are not suitable for networks with both positive and negative links.

- **Dynamic Networks.** Networks are temporally dynamic by nature. There is however a few existing methods for overlapping clustering in dynamic networks (e.g., [12, 15]). These methods are however computationally expensive hence cannot scale to large networks. They also do not allow incremental computation, hence are not suitable for data come in streams.

- **Scalability.** Earlier works on mining overlapping clusters are mostly based on some dense sub-network definitions (e.g., [14, 5]). They therefore develop some brute force search methods for such subnetworks, and these methods are not designed for large networks. Later, researchers proposed to construct clusters by expanding "local clusters" of nodes starting from some cluster seeds (e.g.,[9, 7]), or using factorization models (e.g., [1, 16, 20, 21]). Despite a significant reduction in complexity, these methods are still not yet scalable as they require exhaustive searches for good seeds or expensive matrix factorization operations. There are also very few state-of-the-art methods that are scalable to large networks [17, 18]. These methods are however based on local optimization which may return poor results due to a bad initialization.

- **Parameter Setting.** Most of the existing methods require some pre-defined parameters, e.g., the number of clusters, and the weight for regularization terms, etc.. The tuning of these parameters is one of the key factors for the methods' performance. Despite a number of overlapping clustering methods have been proposed, the automatic tuning of their pre-defined parameters are still a challenge [19].

Within this context, we offer the following topics for master thesis.

- Overlapping clustering in dynamic weighted networks: For this topic, we aim to propose novel methods for overlapping clustering in dynamic weighted networks. We would also like to propose the methods that allow incremental computation so that to work with data streams.

- Empirical analysis on networks with overlapping cluster structure: For this topic, we aim to investigate insights from networks that reveal or correlate with their cluster structure. These insights are very useful for automatic tuning of the pre-defined parameters in the overlapping clustering methods.

**Requirement.** You should be:

- a self motivated learner

- experienced with programming languages (ideally C++ or Java)

- knowledgeable about basic machine learning models

**Contact.** Interested students are encouraged to email to Mr. Tuan-Anh Hoang at hoang(at)l3s(dot)de for scheduling a meeting.

# Bibliography


[1] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. C. Xing. Mixed membership stochastic blockmodels. *JMLR*, 9(9), 2008.

[2] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *KDD*, 2006.

[3] A. Barrat, M. Barthelemy, R. Pastor-Satorras, and A. Vespignani. The architecture of complex weighted networks. *PNAS*, 2004.

[4] D. Easley and J. Kleinberg. *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press, 2010.

[5] T. Evans and R. Lambiotte. Line graphs, link partitions, and overlapping communities. *Physical Review E*, 80(1):016105, 2009.

[6] I. Farkas, D. Ábel, G. Palla, and T. Vicsek. Weighted network modules. *New Journal of Physics*, 2007.

[7] F. Havemann, M. Heinz, A. Struck, and J. Gläser. Identification of overlapping communities and their hierarchy by locally calculating community-changing resolution levels. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(01):P01023, 2011.

[8] J. Kunegis, J. Preusse, and F. Schwagereit. What is the added value of negative links in online social networks? In *WWW*, 2013.

[9] A. Lancichinetti, S. Fortunato, and J. Kertész. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 11(3):033015, 2009.

[10] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Signed networks in social media. In *CHI*, 2010.

[11] J. Leskovec, D. P. Huttenlocher, and J. M. Kleinberg. Governance in social media: A case study of the wikipedia promotion process. In *ICWSM*, 2010.

[12] Y.-R. Lin, Y. Chi, S. Zhu, H. Sundaram, and B. L. Tseng. Facetnet: a framework for analyzing communities and their evolutions in dynamic networks. In *Proceedings of the 17th international conference on World Wide Web*, pages 685–694. ACM, 2008.

[13] T. Nepusz, H. Yu, and A. Paccanaro. Detecting overlapping protein complexes in protein-protein interaction networks. *Nature methods*, 9(5):471–472, 2012.

[14] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005.

[15] L. Tang, H. Liu, J. Zhang, and Z. Nazeri. Community evolution in dynamic multi-mode networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 677–685. ACM, 2008.

[16] F. Wang, T. Li, X. Wang, S. Zhu, and C. Ding. Community discovery using nonnegative matrix factorization. *Data Min. Knowl. Discov.*, 22(3):493–521, May 2011.

[17] J. J. Whang, D. F. Gleich, and I. S. Dhillon. Non-exhaustive, overlapping $k$-means. In *SDM*, 2015.

[18] J. J. Whang, D. F. Gleich, and I. S. Dhillon. Overlapping community detection using neighborhood-inflated seed expansion. *TKDE*, 28(5):1272–1284, May 2016.

[19] J. Xie, S. Kelley, and B. K. Szymanski. Overlapping community detection in networks: The state-of-the-art and comparative study. *Acm computing surveys (csur)*, 45(4):43, 2013.

[20] J. Yang and J. Leskovec. Community-affiliation graph model for overlapping network community detection. In *ICDM*, 2012.

[21] J. Yang and J. Leskovec. Overlapping community detection at scale: a nonnegative matrix factorization approach. In *WSDM*, 2013.

[22] J. Yang and J. Leskovec. Structure and overlaps of ground-truth communities in networks. *TIST*, 2014.