

Sujet	Towards a Comprehensive Evaluation Framework for Dynamic Community Detection Algorithms
Direction de thèse	Bothorel Cécile
Co-encadrement	Brisson Laurent
Environnement collaboratif national, international, co-tutelle	
Financement visé 1 (50% ou 100%)	ARED
Financement 2 (si 50/50%)	Carnot ou Brest Metropole

Résumé du projet de thèse (10 lignes max)

Research into community structures in dynamic networks is crucial for understanding large, complex real-world networks such as social networks, communication networks, biological networks, or any other graph-modeling problem. There is a wide variety of approaches to dynamic community detection.

This thesis, based on an exhaustive empirical study, will address the major challenge of creating a reference framework for testing and evaluating existing dynamic community detection algorithms. The first objective is to propose a hybrid algorithm combining several existing heterogeneous algorithms. The second objective is to define measures for evaluating algorithmic performance, on the one hand, but also to characterize and classify the different forms of dynamic communities detected by different algorithms. The third objective is to develop a reference benchmark for comparing these algorithms, including real and synthetic adjustable data.

Finally, this thesis will contribute to a more global reflection on the definition of the notion of community, in order to propose breakthrough approaches that are probably plural and intrinsically dynamic.

Description de la thèse : - Sujet détaillé (2 pages max)**Context**

The study of community structure in networks that change over time is an important area of research because many real-world networks, such as social networks, communication networks, and biological networks, are dynamic and evolve over time. Understanding how communities change and evolve in these networks can provide insights into the underlying mechanisms that drive their behavior.

Various algorithms have been developed to tackle the challenge of detecting evolving communities in dynamic networks. These algorithms can be categorized into three groups based on their treatment of historical data [13]. First, instant-optimal community detection decomposes time into periods and applies static algorithms independently to each period. Second, temporal trade-off community detection mitigates instability by discovering communities from previous network states. Lastly, cross-time community detection considers the entire network evolution and employs static methods on an aggregated graph.

Comparing these methods is complex because, similar to static networks, there is no universally accepted definition of what constitutes a community [2]. Community detection algorithms may prioritize optimizing certain metrics, making it challenging to determine what they can detect within a given context.

Furthermore, establishing a benchmark framework for evaluating dynamic community detection algorithms is a significant challenge [3]. While benchmark like LFR adaptation to temporal networks [6], RDyn [12], and Mosaic [1] have been proposed for synthetic temporal networks, these approaches come with their limitations. Choices such as modeling in the form of snapshots, the use of decay metrics, and a preference for non-overlapping and continuous communities inherently exclude certain algorithms that may not align well with these criteria.

Finally, to date, there is currently no established model for community evolution, though efforts have been made to extend the stochastic block model to dynamic networks [11]. Limited research has focused on metrics describing temporal communities, and there is a lack of studies categorizing the typology of temporal community structures. While some works have endeavored to characterize how communities interact over time [9, 10], unlike research in static communities [4, 7, 8], exploring the classification of temporal community structures' typology is a research gap yet to be addressed.

Challenges

This PhD thesis will contribute to the broader consideration of defining the concept of dynamic communities. We plan an exhaustive empirical analysis of the communities revealed by existing algorithms on a large variety of datasets (similar methodology already used for static networks in [5]).

A challenging task is to build a typology of the different forms of structures that can be retrieved at a meso scale and evolving over time by algorithms. Due to the diversity of techniques and the

variety of definitions of what a community is, it would be necessary to define quantitative and qualitative metrics that characterize dynamic partitions and the nature of communities retrieved by algorithms.

A second challenge will be to contribute to the development of a benchmarking framework. This framework will facilitate algorithm comparison by generating synthetic datasets with ground truth data and adjustable parameters that reflects community dynamics diversity. The work will build on the Dyn-Bench benchmark project within the DECIDE team.

Research program

1. Begin with a comprehensive review of existing algorithms for temporal clustering. Identify relevant datasets (with or without ground truth), benchmarking frameworks, and metrics that describe network structures and their evolution. Define distinct classes of algorithms.
2. Propose ourselves a hybrid algorithm combining several existing heterogeneous algorithms
3. Develop an experimental protocol for testing these algorithms, ensuring that it accounts for various types of temporal network dynamics.
4. Conduct an in-depth analysis of dynamic partitions, and dynamic communities, focusing on their structural and dynamic aspects. Propose suitable metrics for this purpose. Compare the typologies of communities generated by algorithms within different contexts. Refine the definition of algorithm classes accordingly.
5. Contribute to the creation of an evaluation framework that includes a synthetic data generator capable of reproducing various community typologies, provide a panel of algorithms and evaluation metrics.

Profile

- Background in graph theory, complex networks and social network analysis, but also in machine learning.
- Computer science skills to contribute to the Python Package Dyn-Bench

Positionnement de la thèse par rapport à des activités de recherche antérieures ou actuelles :

Consolider les recherches dans le domaine de complex networks, pour lequel les aspects temporels sont encore très peu explorés, notamment pour la tâche de partitionnement (clustering). Enrichissement du framework d'analyse de communautés dynamiques Dyn-Bench (benchmark) de l'équipe Decide du Lab-Sticc.

Proposer des outils d'aide à la décision de choix d'algorithmes, positionnant l'équipe comme référence dans le clustering de graphe complexe et dynamique, permettant des applications dans les réseaux sociaux, les réseaux de transport, la cybersécurité ou tout type de terrain applicatif mobilisant l'interaction entre des entités.

References

- [1] Asgari, Y. Cazabet, R., Borgnat, P. (2023) Mosaic benchmark networks: Modular link streams for testing dynamic community detection algorithms. *Complex Networks And Applications 2023*, Nov 2023, Menton, France.
- [2] Aynaud, T., Fleury, E., Guillaume, J. L., & Wang, Q. (2013). Communities in evolving networks: definitions, detection, and analysis techniques. In *Dynamics on and of complex networks, volume 2: applications to time-varying dynamical systems* (pp. 159-200). New York, NY: Springer New York.
- [3] Cherifi, H., Palla, G., Szymanski, B. K., & Lu, X. (2019). On community structure in complex networks: challenges and opportunities. In *Applied Network Science* (Vol. 4, Issue 1). Springer. <https://doi.org/10.1007/s41109-019-0238-9>
- [4] Dao, V. (2018) Characterizing community detection algorithms and detected modules in large-scale complex networks. *Data Structures and Algorithms [cs.DS]*. Ecole nationale supérieure Mines-Télécom Atlantique.
- [5] Dao, V., Bothorel, C., & Lenca, P. (2020). Community structure: A comparative evaluation of community detection methods. *Network Science*, 8(1), 1-41.
- [6] Greene, D., Doyle, D., & Cunningham, P. (2010). Tracking the evolution of communities in dynamic social networks. In *2010 international conference on advances in social networks analysis and mining*. IEEE, 176–183
- [7] Lancichinetti, A., Saramäki, J., Kivela, M., & Fortunato, S. (2010). Characterizing the community structure of complex networks. *PLoS ONE*, 5(8). <https://doi.org/10.1371/journal.pone.0011976>
- [8] Leskovec, J., Lang, K. J., Dasgupta, A., & Mahoney, M. W. (2008). Statistical properties of community structure in large social and information networks. *Proceedings of the 17th International Conference on World Wide Web*, 695–704. <https://doi.org/10.1145/1367497.1367591>
- [9] Mohammadmosaferi, K., & Naderi, H. (2020). Evolution of communities in dynamic social networks: An efficient map-based approach. *Expert Systems with Applications*, 147. <https://doi.org/10.1016/j.eswa.2020.113221>
- [10] Palla, G., Barabási, A.-L., & Vicsek, T. (2007). Quantifying social group evolution. *Nature*, 446(7136), 664–667. <https://doi.org/10.1038/nature05670>
- [11] Pensky, M. and Zhang, T. (2019). Spectral clustering in the dynamic stochastic block model, *Electronic Journal of Statistics, Electron. J. Statist.* 13(1), 678-709.
- [12] Rossetti, G. (2017). Rdyn: Graph benchmark handling community dynamics. *Journal of Complex Networks* 5, 6 (07 2017), 893–912. <https://doi.org/10.1093/comnet/cnx016>
- [13] Rossetti, G. and Cazabet, R. (2018). Community discovery in dynamic networks: A survey. *ACM Comput. Surv.*, vol. 51, no. 2, pp. 1–37.

Contexte, verrous et positionnement

L'étude de la structure des communautés dans les réseaux qui évoluent dans le temps est un domaine de recherche important car de nombreux réseaux du monde réel, tels que les réseaux sociaux, les réseaux de communication et les réseaux biologiques, sont dynamiques et évoluent dans le temps. Comprendre comment les communautés changent et évoluent dans ces réseaux peut nous éclairer sur les mécanismes sous-jacents qui régissent leur comportement.

Divers algorithmes ont été mis au point pour relever le défi de la détection des communautés évolutives dans les réseaux dynamiques. Ces algorithmes peuvent être classés en trois catégories en fonction de leur traitement des données historiques [13]. Premièrement, la détection instantanée optimale des communautés décompose le temps en périodes et applique des algorithmes statiques indépendamment à chaque période. Deuxièmement, la détection de communautés par compromis temporel atténue l'instabilité en découvrant des communautés à partir d'états antérieurs du réseau. Enfin, la détection de communautés transtemporelles prend en compte l'ensemble de l'évolution du réseau et utilise des méthodes statiques sur un graphe agrégé.

La comparaison de ces méthodes est complexe car, à l'instar des réseaux statiques, il n'existe pas de définition universellement acceptée de ce qui constitue une communauté [2]. Les algorithmes de détection de communautés peuvent donner la priorité à l'optimisation de certaines mesures, ce qui rend difficile la détermination de ce qu'ils peuvent détecter dans un contexte donné.

En outre, l'établissement d'un cadre de référence pour l'évaluation des algorithmes de détection de communautés dynamiques constitue un défi de taille [3]. Bien que des benchmarks de référence tels que l'adaptation de LFR aux réseaux temporels [6], RDyn [12] et Mosaic [1] aient été proposés pour les réseaux temporels synthétiques, ces approches ont leurs limites. Des choix tels que la modélisation sous forme d'instantanés, l'utilisation de métriques de décroissance et une préférence pour des communautés continues et sans chevauchement excluent de manière inhérente certains algorithmes qui pourraient ne pas s'aligner sur ces critères.

Enfin, il n'existe à ce jour aucun modèle établi pour l'évolution des communautés, bien que des efforts aient été faits pour étendre le modèle stochastique par blocs aux réseaux dynamiques [11]. Peu de recherches se sont concentrées sur les métriques décrivant les communautés temporelles, et il y a un manque d'études catégorisant la typologie des structures des communautés temporelles. Bien que certains travaux se soient efforcés de caractériser la manière dont les communautés interagissent au fil du temps [9, 10], contrairement à la recherche sur les communautés statiques [4, 7, 8], l'exploration de la classification de la typologie des structures communautaires temporelles est une lacune de la recherche qui n'a pas encore été comblée.

Or, comprendre la nature des communautés produites par les algorithmes, comprendre les mesures et proposer une typologie claire de choix d'algorithmes est une nécessité pour tout décideur/utilisateur qui fouille des données d'interactions, que ce soit dans un cadre de réseau de transport, d'énergie ou social.

Programme de recherche

1. Commencer par un état de l'art en matière d'algorithmes pour le clustering temporel de réseaux. Identifier les ensembles de données pertinents (avec ou sans vérité terrain), les cadres d'évaluation comparative et les mesures qui décrivent les structures des réseaux et leur évolution. Définir des classes distinctes d'algorithmes.
2. Proposer un ou plusieurs algorithmes, par exemple hybrides combinant plusieurs algorithmes hétérogènes existants.
3. Élaborer un protocole expérimental pour tester ces algorithmes, en veillant à ce qu'il tienne compte des différents types de la dynamique temporelle des réseaux.
4. Mener une analyse approfondie des partitions dynamiques et des communautés dynamiques, en se concentrant sur leurs aspects structurels et dynamiques. Proposer des mesures appropriées à cette fin. Comparer les typologies de communautés générées par les algorithmes dans différents contextes. Affiner la définition des classes d'algorithmes en conséquence et les conditions d'usage préconisées.
5. Contribuer à la création d'un cadre d'évaluation comprenant un générateur de données synthétiques capable de reproduire diverses typologies de communautés, fournir un panel d'algorithmes et des mesures d'évaluation.