THESE DE DOCTORAT DE L'ETABLISSEMENT UNIVERSITE BOURGOGNE FRANCHE-COMTE

PREPAREE A L'UNIVERSITE DE BOURGOGNE

Ecole doctorale n°37

Sciences Physiques pour l'Ingénieur et Microtechniques (SPIM)

Doctorat en Informatique

Par

Mme. Stephany RAJEH

Impact de la structure communautaire sur la dynamique des réseaux complexes

Thèse présentée et soutenue à Dijon, le 26/5/2023

Composition du Jury :

M. Matthieu LATAPY	Directeur de recherche, Sorbonne Université / CNRS, France	Président (Examinateur)
Mme. Sabrina GAITO	Professeure, Université de Milan, Italie	Rapportrice
Mme. Maria MALEK	MCF-HDR, Cergy Paris Université, France	Rapportrice
Mme. Hamida SEBA	MCF-HDR, Université Claude Bernard, France	Rapportrice
Mme. Sara NAJEM	MCF, Université Américaine de Beyrouth, Liban	Examinatrice
M. Luis M. ROCHA	Professeur, Université d'État de Binghamton à New York, États-Unis	Examinateur
M. Hocine CHERIFI	Professeur, Université de Bourgogne, France	Directeur de thèse
Mme. Marinette SAVONNET	MCF-HDR, Université de Bourgogne, France	Codirectrice de thèse

ACKNOWLEDGEMENT

And know that I am with you always, to the end of age.

Jesus Christ
Matthew 28:20

I want to express my indebtedness to my thesis supervisor, Hocine Cherifi, for his limitless generosity in providing knowledge and expertise throughout all my journey. His devoted involvement is one of the pillars of becoming who I am now. I will forever be grateful for his presence as a mentor. I would also like to express my gratitude to Marinette Savonnet, my co-supervisor. She has always been there when I needed her guidance.

I wish to express my profound gratitude to my reviewers, Sabrina Gaito, Professor at University of Milan, Maria Malek, Associate Professor at CY Cergy Paris University, and Hamida Seba, Associate Professor at University Claude Bernard of Lyon, for accepting to be the reviewers of my thesis. I am grateful for their evaluation, given their expertise and insightful comments that are crucial in shaping my work. Thank you for your time, dedication, and valuable contributions.

I also want to thank Matthieu Latapy, Research Director at the CNRS and Sorbonne University, Sara Najem, Assistant Professor at the American University of Beirut, and Luis M. Rocha, Professor at University Binghamton State University of New York, for accepting to be a part of my defense jury. I am truly honored to have you as part of my thesis.

To my mother, Rita, my father, Tony, and my sister, Jennifer, nothing would have been possible without your existence and constant love. This one is for you.

In memory of my grandmother, Milia, who gave me infinite love since the day I opened my eyes, in memory of my grandfather, Michael, who believed in me and gave me advices to carry on throughout all my life, and in memory of my best friend, Maria, who always stood by me in the bad days before the good ones. This one is for you.

To my friends, Ahmad, Carine, Sarah, and Yasmina, who have always been there with unfailing support throughout my Ph.D. journey. This one is for you.

ABSTRACT

Networks are everywhere. We encounter them daily in our lives, through our social interactions, how we come up with decisions in our brain, to having phone calls, conducting financial transactions, and traveling from one place to another. Individual actions are influenced by their environment, which is, in turn, influenced by the network's topology. Notably, individuals may change their actions, ideas, or opinions to conform to the aspirations of a particular social group. In the same vein, the spread of a virus can take a certain course if the network's structure induces specific pathways for expansion. In such scenarios, communities substantially impact the evolution of the dynamics. They can hinder or enhance diffusion flow depending on where diffusion originates. Nodes within and between communities are responsible for initiating the dynamic diffusion flow in networks, while influential nodes can play a crucial role in boosting diffusion. The significance of comprehending the community structure of a network and its impact on the underlying dynamics, initiated by the nodes, is accentuated by many real-world scenarios. In this thesis, we study the interplay between dynamic models, influential nodes, the process of identifying them, and the network's topology. First, we investigate how the output of various dynamic models is influenced by the network topology, with seed nodes ranked using community-aware centrality measures. Studying this problem can provide insights into how diffusion spreads and identify constraints that limit the effectiveness of utilizing dynamic scenarios in practical situations, such as promoting viral marketing or combating false information. Second, we tackle the problem of influence redundancy and propose a new ranking scheme to naturally select distant nodes to expand any diffusion phenomena. By tackling this problem through the proposed ranking scheme, diffusion ought to be maximized, independent of the network type. This renders a powerful tool suitable for researchers aiming to maximize diffusion in many applications. Third, researchers mainly focus on identifying influential nodes in networks with a non-overlapping community structure, while many networks have an overlapping community structure. Moreover, the measures developed for networks with an overlapping community structure are inflexible to missing or varying information. Therefore, we propose a flexible framework that identifies influential nodes in networks with incomplete, complete, fuzzy, or crisp overlapping information about the nodes. This framework allows researchers to incorporate various information about overlaps and customize it to different circumstances and information availability.

Keywords: Network science, Complex networks, Community structure, Dynamic models, Centrality measures, Information diffusion

RÉSUMÉ

Les réseaux sont partout. Nous les rencontrons quotidiennement dans nos vies, à travers nos interactions sociales, la façon dont nous prenons des décisions dans notre cerveau, les appels téléphoniques, les transactions financières et les déplacements d'un endroit à un autre. Les actions individuelles sont influencées par leur environnement, qui est, à son tour, influencé par la topologie du réseau. Notamment, les individus peuvent modifier leurs actions, leurs idées ou leurs opinions pour se conformer aux aspirations d'un groupe social particulier. Dans le même ordre d'idées, la propagation d'un virus peut prendre un certain cours si la structure du réseau induit des voies d'expansion spécifiques. Dans de tels scénarios, les communautés ont un impact substantiel sur l'évolution de la dynamique. Ils peuvent entraver ou améliorer le flux de diffusion selon l'origine de la diffusion. Les nœuds au sein et entre les communautés sont chargés d'initier le flux de diffusion dynamique dans les réseaux, tandis que les nœuds influents peuvent jouer un rôle crucial dans la stimulation de la diffusion. L'importance de comprendre la structure communautaire d'un réseau et son impact sur la dynamique sous-jacente, initiée par les nœuds, est accentuée par de nombreux scénarios du monde réel. Dans cette thèse, nous étudions l'interaction entre les modèles dynamiques, les nœuds influents, leur processus d'identification et la topologie du réseau. Tout d'abord, nous étudions comment la sortie de divers modèles dynamiques est influencée par la topologie du réseau, avec des nœuds de départ classés à l'aide de mesures de centralité communautaire. L'étude de ce problème peut fournir des informations sur la façon dont la diffusion se propage et identifier les contraintes qui limitent l'efficacité de l'utilisation de scénarios dynamiques dans des situations pratiques, telles que la promotion du marketing viral ou la lutte contre les fausses informations. Deuxièmement, nous abordons le problème de la redondance d'influence et proposons un nouveau schéma de classement pour sélectionner naturellement les nœuds distants afin d'étendre tout phénomène de diffusion. En abordant ce problème à travers le schéma de classement proposé, la diffusion doit être maximisée, indépendamment du type de réseau. Cela rend un outil puissant adapté aux chercheurs visant à maximiser la diffusion dans de nombreuses applications. Troisièmement, les chercheurs se concentrent principalement sur l'identification des nœuds influents dans les réseaux avec une structure communautaire sans chevauchement, alors que de nombreux réseaux ont une structure communautaire qui se chevauche. De plus, les mesures développées pour les réseaux dont la structure communautaire se chevauche sont inflexibles face aux informations manquantes ou variables. Par conséquent, nous proposons un framework flexible qui identifie les nœuds influents dans les réseaux avec des informations incomplètes, complètes, floues ou nettes qui se chevauchent sur les nœuds. Ce framework permet aux chercheurs d'intégrer diverses informations sur les chevauchements et de les personnaliser en fonction des circonstances et de la disponibilité des informations.

Keywords: Science des réseaux, Réseaux complexes, Structure communautaire, Modèles dynamiques, Mesures de centralité, Diffusion de l'information

Contents

1

INTRODUCTION

Contents

1.1/ CONTEXT AND PROBLEMATIC

Diverse and highly connected networks surround us at any given moment in time. Even though networks originate from a multitude of domains, such as biology, finance, transportation, and society, they tend to share similar structural features. One of the most natural and pervasive features is the community structure: networks are often composed of nodes grouped into communities that are more densely connected than nodes in other communities [1]. Communities can correspond to a group of proteins interacting for proper cell functioning [2], to web pages of similar topics [3], to papers related to a specific topic [4], or to a group of friends sharing similar interests and aspirations either in real life or online [5, 6].

Within these networks, it is of utmost importance to identify influential nodes. For instance, in marketing and awareness campaigns, the person's adoption of new products, services, and opinions relies heavily on influencer(s) identification. Pinching essential proteins is decisive for treating infectious and cancer diseases in biological networks. In finance, detecting key financial institutions that may default is necessary to avoid a cascade of defaults. Vaccinating influential nodes allows controlling an epidemic spreading, saving populations from significant threats. Even in terrorist networks, identifying influential nodes is a critical proactive measure to eradicate life losses.

Given such real-world scenarios, one can employ centrality measures, which are considered to be one of the most popular approaches exploiting network structure to identify influential nodes [7]. Nodes with a high centrality value are deemed important. The notion of importance depends on how the centrality functions. Local centrality measures are established upon neighborhood information, such as degree centrality. The latter quantifies node influence based on the total number of links a node possesses. Global centrality measures quantify the centrality of a node by inspecting its position in the entire network, such as betweenness centrality. The latter computes the shortest paths of all

possible node pairs in order to measure the extent of each node lying in the computed shortest paths. Both local and global information can also be combined to quantify node importance [8, 9].

Diffusion is a crucial recurring scenario in many real-world networks. Epidemics, computer viruses, information (whether genuine or fake), opinions, behaviors, and innovations can all diffuse in networks. Diffusion can create adverse outcomes such as pandemics, losing internet connectivity across regions, and spreading fake news. However, it can also yield positive outcomes such as identifying cyber-attackers and the origins of fake news, raising people's awareness, spreading positivity, and conducting profitable marketing campaigns. These scenarios led researchers to introduce models to characterize diffusion better, aiming to either minimize the negative outcomes or maximize the positive ones [10].

Based on the hypothesis that central nodes (i.e., nodes with high centrality) should spread the most [11], the selection of the seeds aims to maximize diffusion in any diffusion phenomenon. For instance, in marketing and awareness campaigns, selecting the most influential nodes to diffuse their influence across the network is decisive as the budget is limited. Indeed, it was shown that in viral marketing, centrality measures based on random walks yield the highest number of consumer activation [12]. By virtue of the node's influence, online word-of-mouth can provide a path for the exponential spread of information [13]. In the same vein, in an epidemic spreading scenario, it was shown that the infectious capacity of nodes is related to their coreness and degree [14].

Notwithstanding their merit, classical centrality measures ignore the network organization in communities, one of the main features characterizing real-world networks [6]. Communities can impact the spreading power of nodes. A diffusion can die out in its originating community if the number of inter-community links (i.e., links connecting communities) does not pass a certain threshold [15, 16]. For a global cascade to occur, networks need to have an optimal fraction of inter-community links [17]. Indeed, communities induce a confinement effect which may trap the spread of any dynamic process [18, 19].

Community-aware centrality measures tackle the shortcoming of classical centrality measures that are agnostic about the community structure [20–32]. They are built on the hypothesis of Granovetter [33], who argued that weak ties could be more powerful than strong ties for diffusion to occur across communities. Weak ties resemble inter-community links (i.e., links between nodes belonging to different communities) and are often less frequent. Strong ties resemble intra-community links (i.e., links between nodes belonging to the same community). Community-aware centrality measures distinguish between these two types of links, which in turn characterize nodes' local and global influence at the mesoscopic level. Intra-community links play a part in the diffusion inside the communities, while inter-community links permit the diffusion to spread across communities.

The hierarchical relationship between the network, the seed nodes, and the diffusion models is shown in Figure 1.1. We use this hierarchy to clarify the main limitations found in the literature and, subsequently, the research questions of the thesis. Given a network (A), the regions in which diffusion begins are dictated by the seeds (B) that initiate a spreading phenomenon based on a diffusion model (C). The seeds can be chosen given any centrality measure, whether classical or community-aware. The network structure impacts centrality measures. For instance, an influential node with a weak community structure strength may not be significant in a network with a strong community structure strength, and vice versa. Following the hierarchical relationship from bottom to top, one

can assess the spreading reachability of the diffusion model (C), such that the extent of the spreading is dependent on the seed nodes which initiate the diffusion (B) and the network structure (A). The selection of the seed nodes using centrality measures (B) also depends on the network structure (A).

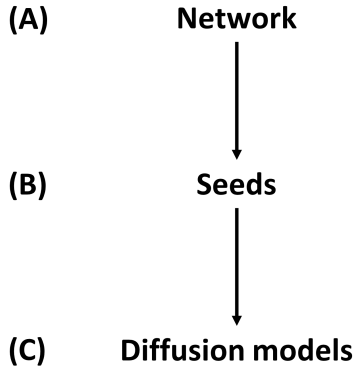


Figure 1.1: **The dependencies between the network, the seed nodes, and the diffusion models.**

Within the realm of the crucial importance of selecting seed nodes that maximize diffusion in networks and that many diffusion models are available in the literature, there are several research problems left unanswered between the network and its underlying structure, the seed nodes, and the diffusion models:

The first research problem: Researchers developed many diffusion models to characterize realistic situations. Each model involves various conditions on how the propagation evolves from one node to another. Given that community-aware centrality measures can exploit community structure information to identify influential nodes, most research focuses on one spreading model to assess their diffusive ability. A single model is not enough to determine the interplay between the diffusive power of seed nodes selected by the community-aware centrality measures, the network structure, and the diffusion model dynamics. In other words, it is unknown how sensitive the different models' spreading reachability is to the network's structure and the seed nodes selected by the community-aware centrality measures.

The second research problem: Since classical centrality measures do not consider the community structure while computing the centrality of each node, they are susceptible to quantifying many influential nodes located in the vicinity of each other. Whether the centrality is based on local information, global information, or both, the top nodes to be targeted for diffusion maximization may all be positioned in one region. Consequently, since the nodes' influence overlaps, their diffusive impact on their direct and indirect neighborhoods is diminished. In other words, these nodes will influence a similar set of nodes while other nodes in distant regions are left intact. Thus, even with a large number of influential nodes to be invested in (i.e., infected or activated), the diffusion spread may be marginal.

The third research problem: Despite the recent advancements in community-aware centrality measures, few works have been dedicated to networks with overlapping community structure. The overlapping community structure can be naturally found in social, collaboration, biological, and ecological networks. Within these networks, a node may pertain to several communities to play several roles. Moreover, the overlapping

community-aware centrality measures do not handle missing or varying overlapping information, which is a challenge faced by many researchers and practitioners. Despite their effectiveness in many real-world scenarios, this leads to underusing the current overlapping community-aware centrality measures to identify influential nodes.

In this thesis, we address the stated research problems and aim to answer the following questions:

1. How does the diffusion models' output depend on the seeds selected by the community-aware centrality measures and the network structure?
2. Can we generalize the rankings of centrality measures using a community-aware ranking scheme to boost diffusion in all network types?
3. Given the vastly changing and diverse, overlapping community-level information, can we effectively identify influential nodes to maximize diffusion?

In the following sections and we describe the contributions of the thesis, its structure, followed by the publications within the framework of the thesis.

1.2/ THESIS CONTRIBUTION

This thesis aims to contribute to the literature on the relationship between network structure, influential nodes identification, and dynamic models. To this end, three main contributions are introduced. These contributions have led to the publication of several scientific articles that we quote in the upcoming sections.

1) The first research study analyzes the relationship between dynamic models, community structure, and community-aware centrality measures. Researchers have designed many dynamic models to characterize real-life spreading scenarios, such as the diffusion of opinions or the spread of epidemic diseases. These models are designed with specific conditions as the dynamic model evolves. We address the question of how the network's community structure can impact the dynamics of four state-of-the-art diffusion models (i.e., Susceptible-Infected (SI), Susceptible-Infected-Recovered (SIR), Linear Threshold (LT), and Independent Cascade (IC)), on a set of synthetic and real-world networks. The diffusive phenomenon starts from a limited set of nodes based on the community-aware centrality measures. Using a set of synthetic and real-world networks with diverse community structure strengths, we highlight the differences and similarities between dynamic models when diffusion originates from various points in the network. The contributions of this study are as follows:

1. Helping better understand how nodes' influence spreads under various diffusion models.
2. Highlighting how the network structure can impact the selection of seed nodes based on community-aware centrality measures and the extent of reachability of any diffusion model.
3. Providing a solid outset for practitioners to select seed nodes that maximize diffusion based on the network structure, budget availability, and the diffusion model that applies in their research case.

2) The second research study is concerned with the problem of influence redundancy. A major pitfall of centrality measures is deeming nodes near each other the most influential. Hence, these nodes are ranked on top of the list in descending order to be selected for maximizing diffusion. However, since these nodes are located near each other, they saturate their shared zone of influence. Consequently, the influence of these nodes has few distinct venues for expansion to distant network regions. To tackle this issue, we propose a ranking strategy exploiting the ubiquity of the community structure in real-world networks. By iterating communities from largest to smallest, the proposed community-aware ranking scheme naturally selects a set of faraway spreaders with the most significant influence. The ranking scheme is tested on synthetic and real-world networks and compared against six classical centrality measures based on the node's neighborhood, paths, and direct and indirect influence in the network. Under all circumstances, the diffusion phenomenon is boosted with the community-aware ranking scheme compared to the descending-order ranking scheme of the centrality measures. The contributions of this study are as follows:

1. Designing a ranking scheme capable of working with any centrality measure computed on any network type (undirected/directed and unweighted/weighted).
2. Assuring the selection of distant influential nodes to expand any diffusion phenomena based on any given budget.
3. Serving as a tool for many real-world cases, such as implementing viral marketing, conducting awareness campaigns, and hindering misinformation on networks.

3) The third research study addresses the issue of identifying influential nodes in complex networks with overlapping community structure. Researchers mainly focus on networks with a non-overlapping community structure. Many networks, such as biological and social, are made up of nodes that pertain to several communities rather than one. Moreover, the overlapping information might not be available for all the nodes or might vary from one node to another. For instance, the node may fully belong to more than one community. However, the node may also belong to more than one community to a specific extent. The dissimilarity in information availability calls for a general approach to identify influential nodes in networks with an overlapping community structure. In this study, we introduce the Overlapping Modularity Vitality framework. The proposed framework can integrate multiple definitions of overlapping modularity via different formulations of the community membership strengths of the nodes. We show that overlapping information provides an exploitable ground for identifying influential nodes effectively with various types of information at hand. The contributions of this study are as follows:

1. Introducing a framework in which one can integrate multiple definitions of overlapping modularity via different formulations of the nodes' community membership strengths (fuzzy or crisp) to identify influential nodes.
2. Investigating how various overlapping modularity alternative definitions that integrate distinct contextual information about the nodes impact identifying influential nodes.
3. Offering flexibility in case missing or varying overlapping information is confronted as the framework can adapt to compute the centrality of one or many nodes with various information types.

1.3/ THESIS STRUCTURE

The structure of the thesis is as follows:

Chapter 1 - Introduction: In this chapter, we discuss the ubiquity of communities, the importance of identifying influential nodes within them, and maximizing diffusion under different dynamic conditions. We shed light on why community-aware centrality measures provide better insights about nodes in networks with community structure and how community structure plays a role in confining diffusion dynamics. Since the network's community structure, influential nodes identification, and dynamic models are linked (dynamics are initiated by nodes that are located in or between communities), in this chapter, we highlight three main research gaps found in the literature, and we pose three main research questions targeting these gaps. The research questions' answers are the basis of the thesis' contributions.

Chapter 2 - Background: In this chapter, we provide the basic notions and elements to grasp the thesis better. We first present the different centrality measures investigated in this thesis: classical and community-aware. The former can be divided into neighborhood-based, path-based, and iterative refinement-based. The latter can be divided into non-overlapping and overlapping. Then, we present the four diffusion models used in this thesis, two of which belong to epidemic models and two to information diffusion models. Finally, we present several network quality measures that could be used to characterize networks while putting more emphasis on overlapping modularity, which is the foundation of the framework proposed in Chapter 5.

Chapter 3 - Diffusion on Networks with Community Structure: In this chapter, we explore the relationship between the network's community structure, diffusion of dynamic models, and nodes selected, given the ranking of community-aware centrality measures. We use four models that portray other dynamical conditions: SI, SIR, LT, and IC. Each model is initiated with a set of seed nodes selected based on the community-aware centrality measures on synthetic and real-world networks. The size of the seed nodes varies according to a predefined budget. We compare and highlight the consistency in the performance of the community-aware centrality measures across one set of models and show how the nodes selected in different regions in the network yield various diffusion outcomes. In summary, the findings of this chapter demystify the performance of the community-aware centrality measures under diverse community structures and diffusion models, which in turn can better guide practitioners in utilizing community-aware centrality measures.

Chapter 4 - A Community-Aware Ranking Scheme: This chapter focuses on the problem of influence redundancy and ranking influential nodes. Centrality measures rank nodes using the classical descending order ranking scheme, with the top being the most influential. However, the top nodes ranked as most influential may be in the vicinity of each other. Consequently, there are diminishes in return for the diffusion initiated by these nodes. To tackle this issue, we propose a community-aware ranking scheme that ranks the most influential nodes by iterating across the communities, from the biggest to the smallest, using any centrality measure. The ranking scheme is assessed with the Susceptible-Infected-Recovered (SIR) diffusion model on a set of synthetic and real-world networks using six centrality measures: two local measures exploiting the neighborhood of the node, two path-based measures, and two iterative refinement-based measures. In summary, the findings of this chapter allow the selection of influential nodes across all

the network regions without saturating their zone of influence and independently of the centrality and network types.

Chapter 5 - The Overlapping Modularity Vitality Framework: This chapter is dedicated to networks with an overlapping community structure, a feature neglected by many researchers working on identifying influential nodes. The researchers who address this issue provide measures that are hard-coded and thus inflexible in many real-world scenarios. To tackle this issue, we propose a flexible framework called Overlapping Modularity Vitality, based on a generalized modularity equation that accounts for the overlapping community structure in a network. Depending on information availability, it can incorporate fuzzy and/or crisp overlapping information for one or many nodes. The framework is assessed with various definitions of the overlapping modularity on the SIR model and is compared with other state-of-the-art overlapping community-aware centrality measures. In summary, the findings of this chapter show that overlapping information, even if it varies from one node to another, can be exploited using the proposed framework to maximize diffusion.

Chapter 6 - Conclusion and Future Perspectives: This chapter concludes our work then suggests pathways for future research work.

1.4/ PUBLICATIONS

JOURNALS

Rajeh, S., & Cherifi, H. (2023). "A community-aware centrality framework based on overlapping modularity." *Social Network Analysis and Mining*, 13(1), 37.

Rajeh, S., Savonnet, M., Leclercq, E., & Cherifi, H. (2023). "Comparative evaluation of community-aware centrality measures." *Quality & Quantity*, 57(2), 1273-1302.

Rajeh, S., & Cherifi, H. (2022). "Ranking influential nodes in complex networks with community structure." *Plos One*, 17(8), e0273610.

Rajeh, S., Savonnet, M., Leclercq, E., & Cherifi, H. (2021). "Characterizing the interactions between classical and community-aware centrality measures in complex networks." *Scientific Reports*, 11(1), 1-15.

Rajeh, S., Savonnet, M., Leclercq, E., & Cherifi, H. (2020). "Interplay between hierarchy and centrality in complex networks." *IEEE Access*, 8, 129717-129742.

CONFERENCES

Zein, H., Yassin, A., Rajeh, S., Jaber, A., & Cherifi, H. (2023). "Community-Aware Centrality Measures Under the Independent Cascade Model." In *International Conference on*

Complex Networks and Their Applications (pp. 588-599). Springer, Cham.

Rajeh, S., Savonnet, M., Leclercq, E., & Cherifi, H. (2022). "Modularity-Based Backbone Extraction in Weighted Complex Networks." In *International Conference on Network Science* (pp. 67-79). Springer, Cham.

Rajeh, S., Yassin, A., Jaber, A., & Cherifi, H. (2021). "Analyzing Community-Aware Centrality Measures Using the Linear Threshold Model." In *International Conference on Complex Networks and Their Applications* (pp. 342-353). Springer, Cham.

Rajeh, S., Savonnet, M., Leclercq, E., & Cherifi, H. (2021). "Comparing community-aware centrality measures in online social networks." In *International Conference on Computational Data and Social Networks* (pp. 279-290). Springer, Cham.

Rajeh, S., Savonnet, M., Leclercq, E., & Cherifi, H. (2021). "Identifying influential nodes using overlapping modularity vitality." In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 257-264).

Rajeh, S., Savonnet, M., Leclercq, E., & Cherifi, H. (2021). "How Correlated Are Community-Aware and Classical Centrality Measures in Complex Networks?." In *International Conference on Complex Networks* (pp. 120-132). Springer, Cham.

Rajeh, S., Savonnet, M., Leclercq, E., & Cherifi, H. (2020). "Investigating centrality measures in social networks with community structure." In *International Conference on Complex Networks and Their Applications* (pp. 211-222). Springer, Cham.

Extended Abstracts

Rajeh, S., Cherifi, H. (2023). "A community-aware ranking scheme to identify influential nodes in complex networks." *International Conference on Complex Networks (CompleNet)*, Aveiro - Portugal.

Rajeh, S., Yassin, A., Jaber, A., & Cherifi, H. (2022). "On Community-aware Centrality Measures Under the Linear Threshold Model." *International School and Conference on Network Science (NetSciX)*, Porto - Portugal.

Rajeh, S., Savonnet, M., Leclercq, E., & Cherifi, H. (2021). "A Comparison of Community-aware Centrality Measures in Online Social Networks." *Conference on Complex Systems (CCS)*, Lyon - France.

Rajeh, S., Savonnet, M., Leclercq, E., & Cherifi, H. (2021). "An empirical study on classical and community-aware centrality measures in complex networks." *Conference on Complex Systems (CCS)*, Lyon - France.

Rajeh, S., Savonnet, M., Leclercq, E., & Cherifi, H. (2021). "Correlation and Topological Analyses of Classical and Community-aware Centrality Measures in Complex Networks." *The 5th European Conference on Social Networks (EUSN)*, Naples - Italy.

Rajeh, S., Savonnet, M., Leclercq, E., & Cherifi, H. (2021). "Spreading Power of Key Nodes in Online Social Networks with Community Structure." *epiDAMIK 2021: 4th epi-*

DAMIK ACM SIGKDD International Workshop on Epidemiology meets Data Mining and Knowledge Discovery, Singapore.

Rajeh, S., Savonnet, M., Leclercq, E., & Cherifi, H. (2021). "Investigating the Relationship Between Community-aware and Classical Centrality Measures." *Networks 2021: A Joint Sunbelt and NetSci Conference*.

Rajeh, S., Savonnet, M., Leclercq, E., & Cherifi, H. (2021). "Comparaisons des mesures de centralité classiques et communautaires : une étude empirique." *Conférence Internationale Francophone sur la Science des Données*, Marseille - France (Invited paper).

Rajeh, S., Savonnet, M., Leclercq, E., & Cherifi, H. (2021). "Classical versus Community-aware Centrality Measures: An Empirical Study." *French Regional Conference on Complex Systems (FRCCS)*, Dijon - France.

Rajeh, S., Savonnet, M., Leclercq, E., & Cherifi, H. (2021). "Analyse des mesures de hiérarchie et de centralité dans les grands graphes de terrain." *Extraction et Gestion des Connaissances: Actes EGC'2021*, Montpellier - France.

Rajeh, S., Savonnet, M., Leclercq, E., & Cherifi, H. (2020). "Assessing the Relationship Between Centrality and Hierarchy in Complex Networks," *The 9th International Conference on Complex Networks and their Applications (CNA)*, Madrid - Spain.

Rajeh, S., Savonnet, M., Leclercq, E., & Cherifi, H. (2020). "An Empirical Comparison of Centrality and Hierarchy Measures in Complex Networks." *Conference on Network Modeling and Analysis*, Montpellier - France.

POSTERS

Rajeh, S., Cherifi, H. (2023). "A Centrality Ranking Strategy in Modular Complex Networks." *International School and Conference on Network Science (NetSci)*, Vienna - Austria.

Rajeh, S., Cherifi, H. (2023). "Identifying Influential Nodes: The Overlapping Modularity Vitality Framework." *French Regional Conference on Complex Systems (FRCCS)*, Le Havre - France.

Rajeh, S. & Cherifi, H. (2023). "Overlapping Modularity Vitality : Une mesure d'influence dans les réseaux complexes à structure communautaire avec recouvrement." *Extraction et Gestion des Connaissances (EGC)*. Lyon - France.

Rajeh, S., Savonnet, M., Leclercq, E., & Cherifi, H. (2022). "A Modular Backbone For Weighted Complex Networks." *NetSci 2022*. Shanghai - China.

Rajeh, S., Savonnet, M., Leclercq, E., & Cherifi, H. (2022). "A Modularity Backbone Extraction Method for Weighted Complex Networks." *IC2S2 2022: 8th International Conference on Computational Social Science*. Chicago - United States.

Rajeh, S., Savonnet, M., Leclercq, E., & Cherifi, H. (2022). "A Modularity Backbone Extraction Method for Weighted Complex Networks." *Fifth Northeast Regional Conference on Complex Systems*. New York - United States.

Rajeh, S., Savonnet, M., Leclercq, E., & Cherifi, H. (2021). "Analyzing the Correlation of Classical and Community-aware Centrality Measures in Complex Networks." *IC2S2 2021: 7th International Conference on Computational Social Science*, Zurich - Swtizerland.

Rajeh, S., Savonnet, M., Leclercq, E., & Cherifi, H. (2021). “Correlation of community-aware and classical centrality measures: Examining the role of network topology.” *Fourth Northeast Regional Conference on Complex Systems*. Binghamton - United States.

Rajeh, S., Savonnet, M., Leclercq, E., & Cherifi, H. (2020). “Hierarchy and Centrality: Two Sides of The Same Coin?.” *Conference on Complex Systems (CCS)*. Thessaloniki - Greece.

2

BACKGROUND

Contents

2.1/ CENTRALITY MEASURES

Centrality measures quantify the node influence in a network. Let $G(V, E)$ be an undirected and unweighted graph where V is the set of nodes, $E \subseteq V \times V$ is the set of edges, and $N = |V|$ is the total size of the network. The connections between the nodes are represented in the adjacency matrix $A = (a_{i,j})$ such that $a_{i,j} = 1$, if node i is connected to node j and $a_{i,j} = 0$, otherwise. Let the neighborhood of any node i be defined as the set $\mathcal{N}_p(i) = \{j \in V, (i, j) \in E\}$ at length p, where $p = 1, 2, ..., D$. D is the diameter of G. Accordingly, two nodes are neighbors of order A^p if there's a minimal path connecting them at p steps.

2.1.1/ CLASSICAL MEASURES

Classically, centrality measures can be categorized into neighborhood-based, path-based, and iterative refinement-based [7] measures. Neighborhood-based centralities count on the node's capacity to influence its surrounding neighborhood. Path-based centralities rely on the node's role in disseminating information quickly and effectively. Finally, iterative refinement-based centralities quantify the importance of a node based on its direct influence on its neighbors, the capacity of the neighbors to influence their neighborhood, and so on. Centrality measures can also be classified into local and global measures depending on the topological information they process. Local measures rely on a

node's ability to influence its neighborhood, while global measures are concerned with the ability of a node to influence the whole network. Generally, local measures require a low computation cost, while global ones are computationally intensive. More recent works consider centrality as a multidimensional issue where local and global information can be combined [8, 9].

NEIGHBORHOOD-BASED MEASURES

1. Degree centrality of a node sums the total number of connections a node has in its direct neighborhood. It can be defined as:

$$\eta_d(i) = k_i = \sum_{j=1}^{N} a_{ij} \tag{2.1}$$

where a_{ij} is obtained from A^1, 1-step neighborhood (p=1).

2. Maximum Neighborhood Component centrality extracts the largest connected component (LCC) from the direct neighborhood of a node to quantify its importance. It can be defined as:

$$\eta_m(i) = |LCC \in \mathcal{N}_1(i)| \tag{2.2}$$

where $\mathcal{N}_1(i)$ is the set of direct neighbors of node i.

PATH-BASED CENTRALITY MEASURES

1. Betweenness centrality is based on the frequency of a node situated in the shortest path between any other two nodes in the network. It is defined as:

$$\eta_b(i) = \sum_{s,t \neq i} \frac{\sigma_i(s,t)}{\sigma(s,t)} \tag{2.3}$$

where $\sigma(s,t)$ is the number of shortest paths between nodes s and t and $\sigma_i(s,t)$ is the number of shortest paths between nodes s and t that pass through node i.

2. Closeness centrality is based on how close, on average, a node is to all other nodes in the network. It is defined as:

$$\eta_c(i) = \frac{N-1}{\sum_{j=1}^{N-1} d(i,j)} \tag{2.4}$$

where $d(i,j)$ is the shortest path distance between node i and j.

ITERATIVE REFINEMENT-BASED CENTRALITY MEASURES

1. Katz centrality quantifies a node's importance based on the influence of all the other nodes on it and their subsequent distances. As the distance of a node increases, its

influence diminishes. It is defined as:

$$\eta_k(i) = \sum_{p=1} \sum_{j=1} s^p a_{ij}^p \tag{2.5}$$

where a_{ij}^p is the connectivity of node i with respect to all the other nodes at A^p and s^p is the attenuation factor where $s \in$ [0,1].

2. PageRank centrality is based on the quantity and quality of the node's direct and indirect connections. It can be thought of as a Markov chain process. It is defined as:

$$\eta_p(i) = \frac{1-d}{N} + d \sum_{j \in \mathcal{N}_1(i)} \frac{\eta_p(j)}{m_j} \tag{2.6}$$

where $\eta_p(i)$ and $\eta_p(j)$ are the PageRank centralities of node i and node j, respectively, $\mathcal{N}_1(i)$ is the set of direct neighbors of node i, m_j is the number of links from node j to node i, and d is the damping parameter where $d \in$ [0,1]. The damping parameter d is set to 0.85.

2.1.2/ COMMUNITY-AWARE MEASURES

Classical centrality measures are community-agnostic. They do not incorporate any information about the community structure to measure the node influence, although it proves to be a ubiquitous property in real-world networks [6, 34]. In contrast, recently developed community-aware centrality measures offer a novel perspective by exploiting the network's mesoscopic properties more effectively to quantify the nodes' influence. They can be divided into non-overlapping [20–27] and overlapping measures [28–32]. In a network with a community structure, one can distinguish the local and global influence of the nodes. A node exerts its local influence on nodes inside its community through its intra-community links. In contrast, its global importance quantifies its ability to influence the nodes it connects to outside its community and is exerted through the node's inter-community links. Non-overlapping community-aware centrality measures combine intra-community and inter-community links differently. Overlapping community-aware centrality measures add overlapping information that further refines the local and/or global influence of a node.

Let graph G be divided to $C = \{c_1, c_2, ..., c_q, ..., c_{|C|}\}$ communities where c_q is q-th community, $|C|$ is the total number of communities, and n_{c_q} is the total number of nodes in community c_q. In a non-overlapping community structure, a node i is a member of a single community c_q, therefore $c_q \cap c_l = \emptyset \ \forall \ q \neq l$ and $\bigcup_{q=1}^{|C|} = V$. In an overlapping community structure, a node i can be a member of one or more communities. Consequently, $\exists \ q \neq l \mid c_q \cap c_l \neq \emptyset$ and $\bigcup_{q=1}^{|C|} = V$. Intra-community edges link nodes in the same community, while inter-community edges join nodes in different communities. More formally, $\mid E_{c_q}^{in} \mid = \frac{1}{2} \sum_{i,j \in c_q} A_{i,j}$ and $\mid E_{c_q}^{out} \mid = \sum_{i \in c_q} \sum_{j \in C \setminus c_q} A_{i,j}$ denote, respectively, the number of intra-community and inter-community edges of community c_q.

A node i has a total degree of $k_i = \sum_{j=1}^{N} A_{i,j} = k_i^{intra} + k_i^{inter}$ where k_i^{intra} is the internal degree and k_i^{inter} is the external degree. More formally, $k_i^{intra} = \sum_{j=1}^{N} A_{i,j} \delta(c_i, c_j)$ and

$k_i^{inter} = \sum_{j=1}^{N} A_{i,j}(1 - \delta(c_i, c_j))$ where $\delta(i,j)$ is the Kronecker delta function, indicating that $\delta(m,n) = 1$ if $m = n$, otherwise $\delta(m,n) = 0$, c_i denotes community of node i, and c_j denotes the community of node j. Moreover, a node i has a degree in community c_q denoted as k_{i,c_q}. In other words, k_{i,c_q} is the number of links node i has, reaching community c_q, defined as $k_{i,c_q} = \sum_{j=1}^{N} A_{i,j}\delta(c_j, c_q)$. It is important to understand that the distinction between k_i^{intra} and k_{i,c_q} lies in the fact that the former represents the overall internal degree of the node across all communities, while the latter refers to the internal degree of the node within a particular community c_q.

NON-OVERLAPPING MEASURES

1. Participation Coefficient [20] quantifies the node's importance based on its participation in various communities through its inter-community links. The more diversified across the communities a node's links are, the higher its Participation Coefficient. If the node has only intra-community links, its Participation Coefficient reduces to zero. It is defined as follows:

$$\alpha_{PC}(i) = 1 - \sum_{q=1}^{|C|} \left(\frac{k_{i,c_q}}{k_i} \right)^2 \tag{2.7}$$

2. Community-based Centrality [21] places importance on the distribution of a node's links in its community and across the other communities. The size of the communities it is connected to is also part of the measure. Indeed, the community size either undermines or enhances the node's influence. It is defined as follows:

$$\alpha_{CBC}(i) = \sum_{q=1}^{|C|} k_{i,c_q} \left(\frac{n_{c_q}}{N} \right) \tag{2.8}$$

3. Comm Centrality [22] differentiates hubs (high-degree nodes) from bridges (the link between communities) based on a weighted combination of the intra-community links and inter-community links while giving bridges a higher priority. It is defined as follows:

$$\alpha_{Comm}(i) = (1 + \mu_{c_q}) \times \left(\frac{k_i^{intra}}{max_{(j \in c_q)} k_j^{intra}} \times R \right) + (1 - \mu_{c_q}) \times \left(\frac{k_i^{inter}}{max_{(j \in c_q)} k_j^{inter}} \times R \right)^2 \tag{2.9}$$

where μ_{c_q} is the fraction of inter-community links over the total community links in the community, and R is a user-defined value to standardize the intra-community and inter-community values.

4. K-shell with Community [23] identifies hubs and bridges depending on their hierarchical position as determined by their k-shell after dividing the network into two components. The first comprises the intra-community links, characterizing the node's local influence. The second comprises the inter-community links, characterizing the node's global influence. Then a weighted linear combination of the two influences is computed to assess the node's importance. It is defined as follows:

$$\alpha_{ks}(i) = \delta \times \alpha^{intra}(i) + (1 - \delta) \times \alpha^{inter}(i) \tag{2.10}$$

where $\alpha^{intra}(i)$ and $\alpha^{inter}(i)$ refer to the k-shell value of node i on the graphs constituting intra-community links and inter-community links, respectively. In this thesis, δ is equal to 0.5 so neither hubs nor bridges are preferentially selected.

5. Community-based Mediator [24] identifies influential nodes that can quickly spread information across communities based on the entropy of their random walks. The more a node connects communities, the higher its entropy and its importance under the Community-based Mediator. It is defined as follows:

$$\alpha_{CBM}(i) = H_i \times \frac{k_i}{\sum_{i=1}^{N} k_i} \tag{2.11}$$

where $H_i = [-\sum \rho_i^{intra} log(\rho_i^{intra})] + [-\sum \rho_i^{inter} log(\rho_i^{inter})]$ is node i's entropy according to ρ^{intra} and ρ^{inter} which represent the intra-community and inter-community links over the total degree of node i and $\sum_{i=1}^{N} k_i$ represents the sum of the degrees of all the nodes.

6. Community Hub-Bridge [25] weighs the node's local influence through its intra-community links by the size of the node's belonging community and the node's global influence by the number of neighboring communities a node can reach in one hop. Then, it sums both influences to assess the overall influence. It is defined as follows:

$$\alpha_{CHB}(i) = n_{c_q,i} \times k_i^{intra} + \sum_{c_l \subset C \setminus c_q}^{N} \bigvee_{j \in c_l} a_{ij} \times k_i^{inter} \tag{2.12}$$

where $n_{c_q,i}$ is the size of the community c_q node i belongs to and $\bigvee_{j \in c_l} a_{ij} = 1$ if node i connects to at least one node j in community c_l.

7. Modularity Vitality [26] identifies hubs and bridges based on their contribution to the network's modularity. One quantifies their contribution through the vitality principle that measures the effect of node removal on a quality measure. Removing hubs tends to decrease the network's modularity, while removing bridges tends to increase it. It is defined as follows:

$$\alpha_{MV}(i) = Q(G) - Q(G \setminus \{i\}) \tag{2.13}$$

where $Q(G)$ is the network's modularity and $Q(G \setminus \{i\})$ is the network's modularity after the removal of node i. Note that since Modularity Vitality is a signed community-aware centrality measure, we investigate it using hubs-first (α_{MV}^{+}), bridges-first (α_{MV}^{-}), and hubs-and-bridges ($|\alpha_{MV}|$) ranking schemes.

8. Map Equation Centrality [27] measures the importance of a node in a network by considering the collective marginal harm it causes to the remaining nodes in terms of codeword length, that is, by how many bits the codeword lengths for the remaining nodes could be reduced if the node was silenced. Silencing a node means that when a random walker visits it, the sender does not communicate the codeword for visiting the node to the receiver, resulting in a compressed network modular description. The more one can compress the network's modular description without encoding the node, the higher the node's influence. This means that nodes frequently visited by the random walker play an important role in the network's modular structure. It is defined as follows:

$$\alpha_{MapEq}(i) = L^i - L^{i*} \tag{2.14}$$

whereL^i denotes the inefficient code (i.e., the difference in the code length between the coding scheme that assigns codewords to all nodes but does not use node i's codeword) and L^{i*} denotes the efficient code (i.e., the coding scheme that assigns codewords to all nodes but never for node i).

Fig 2.1 shows the top node selected by the various community-aware centrality measures on a toy example network with two communities. The top node appears to be either a hub-like or bridge-like node. In this figure, if a community-aware centrality measure is labeled twice, it is because two nodes obtain the same highest centrality score value. In this case, they are both represented as the top node for the given centrality measure. For instance, one can see that hub-like nodes (i.e., nodes 4 and 13) are the top node for Modularity Vitality targeting hubs (α^{+}_{MV}), Modularity Vitality targeting hubs and bridges ($|\alpha_{MV}|$), Community-based Centrality (α_{CBC}), and the Map Equation Centrality (α_{MapEq}). One can expect this result, as these community-aware centrality measures implicitly emphasize hub-like nodes.

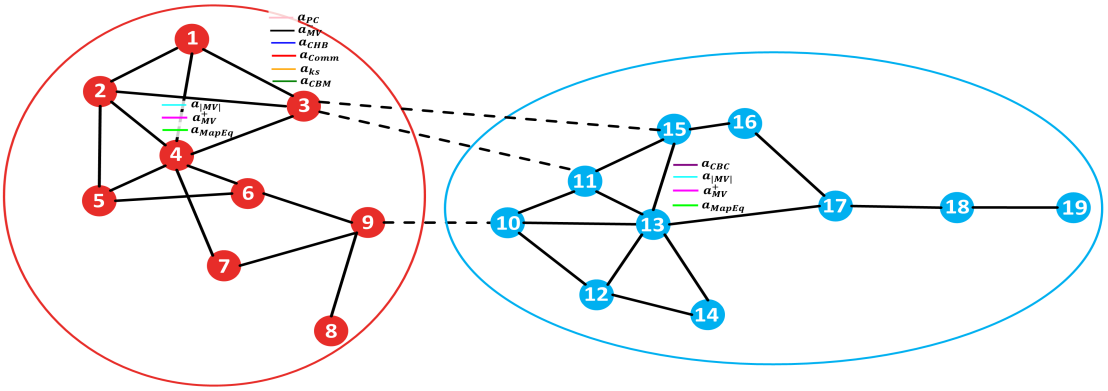


Figure 2.1: **Top nodes selected by the community-aware centrality measures under study.** The 19 nodes network with 31 edges has two communities with $< k >$ = 3.26. The solid lines represent the intra-community links and the dashed lines represent the inter-community links.

Modularity Vitality targeting hubs assesses the node's importance by quantifying the Hubs' contribution to the overall modularity of the network. Removing node 13 or node 4 is the most disruptive action against the cohesiveness of the network communities. Modularity Vitality targeting hubs and bridges takes the aggregate contribution of both hubs and bridges and ranks the nodes accordingly. Both nodes 4 and 13 have the highest effect on the network's modularity compared to the bridge-like nodes. Community-based Centrality chooses node 13 as the top node because it has the highest number of connections in its community. Moreover, it belongs to the largest community. Community-based Centrality prioritizes hub-like nodes located in large communities. The Map Equation Centrality picks nodes 4 and 13 as top nodes (i.e., they have the same centrality value). One can see that they are distributed across the two communities. The random walker is more likely to stay in nodes 4 and 13 as they are mainly responsible for the flow of information inside their communities.

The remaining centrality measures rank node 3 as the top node. It plays a bridge-like role between the communities. Indeed, it is the node with the highest number of inter-community links. Participation Coefficient (α_{PC}) selects it as the top node because it distributes its connections across the two communities. Similarly, the Community-based Mediator (α_{CBM}) grants node 3 the highest entropy based on the distribution of its intra-community and inter-community links. For Comm Centrality (α_{Comm}) and Community Hub-Bridge (α_{CHB}), node 3 is the most critical bridge-like node due to its connections to its external community while playing an essential role inside its community as well. Concerning

K-shell with Community (α_{ks}), node 3 is the mostly embedded node inside the network's local and global components, yielding the highest K-shell with Community centrality value when both influences are linearly combined. Finally, Modularity Vitality targeting bridges (α^{-}_{MV}) ranks node 3 as the top node since it is the node that essentially plays a positive role in the community structure cohesiveness. In other words, if one removes it, the network's modularity increases since it becomes more tightly connected.

OVERLAPPING MEASURES

This section presents the main overlapping community-aware centrality measures. Note that throughout the thesis, when we refer to community-aware centrality measures, we refer to the non-overlapping measures unless stated otherwise.

1. Membership [28] sums up the number of communities a node is part of. Nodes belonging to many communities are considered the most influential, as they play a prime role in propagating diffusion across many communities. The abbreviation of Membership is β_M hereafter.

2. OverlapNeighborhood [29] assumes that overlapping nodes have a higher probability of connecting to different hubs throughout the network. Hence, by randomly targeting the neighbors of overlapping nodes for diffusion or immunization, results are improved compared to solely targeting overlapping nodes. The abbreviation of OverlapNeighborhood is β_{ON} hereafter.

3. Random Walk Overlap Selection [30] proposes to target overlapping nodes based on a random walk since it considers that not all overlapping nodes are important. However, an overlapping node selected by a random walker is more important since the random walker had a higher probability of visiting it due to its high degree rather visiting another overlapping node. The abbreviation of Random Walk Overlap Selection is β_{RWOS} hereafter.

4. Overlapping Modular Centrality [31] first checks if a node is overlapping or not. If so, the node's local influence will be based on the network made up of the intra-community links a node has (i.e., local component). These intra-community links pertain to more than one community. If the node is not overlapping, the node's local influence will still be based on the network made up of intra-community links. However, the node's local influence is less amplified since the node does not participate in several communities. On the other hand, the node's global influence is always the same, whether the node is overlapping or not. The global influence is based on the network constructed by the inter-community links of the node (i.e., global component). Then, choosing any centrality measure, one computes it on the local and global components. The result will be a vector of two values, indicating the local and global influence. Both influences are then combined depending on the user's choice. In this thesis, we use the degree centrality with the modulus combination. The abbreviation of Overlapping Modular Centrality is β_{OMC} hereafter.

2.2/ DIFFUSION MODELS

Diffusion in complex networks is an important interdisciplinary research area representing many real-world situations. Researchers from various domains were attracted to developing models for a more realistic characterization of dynamics on networks. The goal is to describe the current dynamic situation better to apply well-informed decisions and predict future trends. For instance, models were proposed to combat malware spreading across computer networks [35], to optimize online marketing campaigns [36], and to forecast COVID-19 at different territorial levels [37]. Thus, it is clear that one model characterizing all real-world situations is insufficient.

Due to the ubiquity of dynamic interactions across networks in many domains, there is a rich taxonomy for diffusion models. Some researchers refer to them as simple and complex contagions [38–41]. The dynamics of a simple contagion designate that a single contact with an active/infected node is enough for successful transmission. With a complex contagion, a node requires an aggregation of connections with its neighborhood for successful communication to take place. Other researchers divide diffusion models into biological/epidemic models and social/information diffusion models [42–45]. Epidemic models characterize the spread of a virus between individuals, with various parameters in place, such as the infection rate and the recovery rate. In information diffusion, the goal is to simulate the influence of one person over others through passing knowledge, ideas, or opinions toward products or controversial topics. Diffusion models also can be divided into explanatory and predictive models [46–48]. In explanatory models, given an ordered sequence of activated nodes, the goal is to backtrack the evolution of the propagation. In predictive models, the aim is to infer the development of the diffusion process from a set of activated nodes. One can further divide predictive models into graph-based and non-graph-based [46].

Regardless of the taxonomy, popular models mainly differ in three main aspects. The first is the number of states a node can acquire. For instance, in the Susceptible-Infected-Recovered (SIR) model [49], a node can be in one of three states. In contrast, in the Susceptible-Infected (SI) model [49], the node can be either susceptible or infected. The second is the frequency of an activated node capable of influencing other nodes. In the Independent Cascades (IC) model [50], an activated node has a single chance of affecting its neighboring nodes. On the contrary, in the Linear Threshold (LT) model [51], more than one possibility of activation is possible. Finally, the third main difference relates to the conditions set on nodes and/or edges. For example, in the SIR model, a constant infection rate is set, while in the IC model, the probability of influencing neighboring nodes can vary. Note that one can use the terms active/infected and inactive/susceptible interchangeably.

We are interested in using various models to study the interplay of the diffusion process and the networks given a set of activated nodes selected based on the community-aware centrality measures. In this thesis, we consider four diffusion models:

- The Susceptible-Infected (SI) model
- The Susceptible-Infected-Recovered (SIR) model
- The Linear Threshold (LT) model
- The Independent Cascade (IC) model

We choose these models for three main reasons: their popularity in the scientific community, their capacity to model realistically diverse diffusion phenomena, and their genericity.

The SI and the SIR models originate from epidemiological modeling, while the LT and IC models originate from information diffusion modeling. Additionally, the SI, SIR, and IC models are simple contagion processes where an active node has a single chance of activating another node. In this case, an inactive node does not rely on collective influence to change its state. A single event from an influential activated node is enough for it to become active. In contrast, in the LT model, the success of a transmission depends on the aggregation of the activations of a node's neighborhood. Finally, all these models are predictive since they all predict the diffusion spread in a network given a set of activated nodes. Fig 2.2 illustrates the main characteristics of the four models. In the following sections, we discuss each model in more detail.

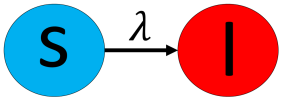
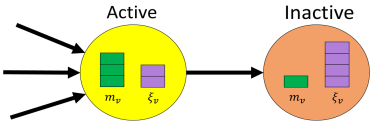
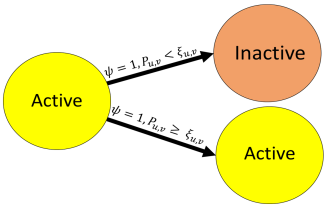
Model	Origin	Type	Contagion	Diffusion Mechanism
SI	Epidemiology	Predictive (non-graph-based)	Simple	
SIR	Epidemiology	Predictive (non-graph-based)	Simple	
LT	Sociology	Predictive (graph-based)	Complex	
IC	Marketing	Predictive (graph-based)	Simple	

Figure 2.2: **Comparing the diffusion models under study.** λ is the infection rate, ψ is the recovery rate, m_v is the total number of active neighbors node v possesses, ξ_v is node the threshold of node v, $P_{u,v}$ is the likelihood of node u activating node v, and $\xi_{u,v}$ is the threshold of edge (u, v).

Note that in this thesis, we use seed-dependent models, not seed-independent ones like the random walk. Thus, if we change the seeds, the output will change. We also note that we are not addressing the problem of influence maximization (i.e., finding the smallest set of nodes that ignites the maximal activation size of nodes). Instead, we are more interested in the interplay between the dynamic models, the network structure, and seed nodes selected using centrality measures with different budget availabilities. The influence maximization problem is NP-hard [52, 53]. Several algorithms have been proposed to approximate this problem. Nevertheless, many require information surpassing the structural-level information. Moreover, many suffer from scalability limitations. This,

in turn, hinders their utilization in many real-world cases. For more information about influence maximization, one can refer to [54, 55].

2.2.1/ EPIDEMIC MODELS

The SI and the SIR models proposed by Kermack and McKendrick [49] are epidemic models characterizing various diseases. They have been initially developed under a "well-mixed" populations hypothesis, where individuals have the same probability of interacting. However, in real-world scenarios, a person's likelihood of contacting another person depends on the underlying contact network structure [42, 56]. Accordingly, scientists integrate the original models on a network, and the status of the nodes evolves according to the contact information [57, 58]. One can distinguish homogeneous mixing (i.e., individuals are equally likely to interact with each other) and heterogeneous mixing (i.e., contact rate depends on each individual). If detailed data about people interactions are missing, one uses a homogeneous mixing approximation [56].

SUSCEPTIBLE-INFECTED (SI) MODEL

In the SI model, a node can be either in susceptible (S) or infected (I) states. For a viral disease, the susceptible state indicates that the node has not yet encountered the virus. An infected node is a node holding the virus. It may infect its susceptible neighbor(s) at time t based on the infection rate λ. Once a node is in the infected state, it remains so forever.

Suppose that at time t, the number of susceptible individuals is denoted by $S(t)$, and the number of infected individuals is denoted as $I(t)$. Concurrently, in a population of $N = S(t) + I(t)$ individuals, the susceptible density is $i(t) = I(t)/N$ and the infected density is $s(t) = S(t)/N$ where $s(t) + i(t) = 1$. Assuming that each individual has, on average, $< k >$ connections, and di/dt represents the rate of change of the number of infected individuals over time, the evolution of the SI model is described as follows:

$$\frac{di}{dt} = \lambda < k > i(t)[1 - i(t)] \tag{2.15}$$

Equation 2.15 indicates that the infected density is directly proportional to the increase in the infection rate and the average number of connections of an individual. Note that since there are only two states, an increase in infected individuals leads to an equal decrease in susceptible individuals.

The SI model is one of the basic models of epidemics, such as several sexually transmitted diseases, yet other significant applications fit into its dynamics. For instance, computer viruses [59], human activity [60], and rumor spreading [61].

SUSCEPTIBLE-INFECTED-RECOVERED (SIR) MODEL

In the SIR model, a node can be either in the susceptible (S), infected (I), or recovered (R) state. Initially, a set of nodes is infected while all the remaining nodes are susceptible. At each time step t, every infected node attempts to infect its neighbor(s) with an infection

rate λ. Simultaneously, each infected node can recover with a recovery rate ψ. It is unable to infect again. Recovered nodes stay in this state throughout the dynamic process.

The population N in the SIR model divides into three subpopulations such that $N = S(t) + I(t) + R(t)$ and the densities are $s(t) = S(t)/N$, $i(t) = I(t)/N$, $r(t) = R(t)/N$, such that $s(t) + i(t) + r(t) = 1$. Assuming that each individual has, on average, $< k >$ connections, and ds/dt, di/dt, and dr/dt represent the change over time in the number of susceptible, infected, and recovered individuals, respectively, then the evolution of the SIR model is described as follows:

$$\frac{ds}{dt} = -\lambda < k > i(t)[1 - r(t) - i(t)] \tag{2.16}$$

$$\frac{di}{dt} = -\psi i(t) + \lambda < k > i(t)[1 - r(t) - i(t)] \tag{2.17}$$

$$\frac{dr}{dt} = \psi i(t) \tag{2.18}$$

Since any individual can recover randomly at any time t, the epidemic outbreak decreases as the evolution of the SIR process proceeds. Note that the SI model can be apprehended as the limiting case of the SIR model with a recovery rate (ψ) equal to zero [56].

Like the SI model, the SIR model is not limited to diseases that yield immunity or death. One can model rumors using the SIR model [62]. Indeed, the SIR model is a basic model for many information diffusion scenarios, such as the SIRaRu model, which simultaneously considers individuals exposed to the rumor and chooses to disseminate it along with individuals who do not [63]. SIR can also be an ideal model for advertising [64], economics [65], and computer viruses [66].

2.2.2/ INFORMATION DIFFUSION MODELS

Information diffusion has recently gained much attention due to the ongoing technological expansion and the rise in social media platforms. Information spreading can take the form of viral marketing [67], innovation [68], norms [69], behaviors [70], fake news [71] and opinions [72]. Modeling diffusion processes helps policymakers and practitioners understand social behaviors, identify terrorists, halt disseminating false information, and improve marketing campaigns compared to traditional approaches [43]. Numerous models have been proposed to describe information diffusion on networks. The most popular and classical ones are the Linear Threshold (LT) model and the Independent Cascade (IC) model [10, 47, 73].

LINEAR THRESHOLD (LT) MODEL

The Linear Threshold (LT) model proposed by Granovetter in 1978 depicts the significant role the neighborhood of an individual plays in social influence [51]. The individual makes a binary choice at every time step t, relying on social reinforcement. In other words, it looks at the opinion of its neighborhood at each time step t and becomes active if and only if the fraction of the activated neighboring nodes (i.e., their collective influence) is greater than a threshold value.

An individual v in the LT model possesses a threshold $\xi_v \in [0,1]$. At any time t, it can be in one of the two alternative states:

$$v(t) = \begin{cases} 0 & \text{if } v \textit{ is inactive} \\ 1 & \text{if } v \textit{ is active} \end{cases} \tag{2.19}$$

At each time step t, an inactive node v checks the states of its neighbors. It is activated if and only if:

$$\frac{m_v}{k_v} \geq \xi_v \tag{2.20}$$

where m_v is the number of active neighbors of node v and k_v is the degree of node v.

Initially developed to characterize complex social contagions, the LT model is prevalent in real-world scenarios such as joining a riot or a strike, diffusing rumors, voting, and residential segregation [51]. It has also been used and refined to predict product adoption in online social networks [74], as well as identifying influential nodes [75], and to assess the robustness of urban railway networks [76].

INDEPENDENT CASCADE (IC) MODEL

Goldenberg *et al.* introduced the Independent Cascade (IC) model to represent marketing dynamics under the simple contagion mechanism [50]. In the IC model, an individual can either be active or inactive. Each activated individual u has only one chance to activate their neighbor v with a probability $P_{u,v}$. The attempt to activate an inactive individual is independent of the remaining activated individuals. In other words, the inactivated individual does not depend on the collective influence of their neighbors. Instead, a single influential node is capable of activating it.

An individual v in the IC model can be only in one of the two alternative states:

$$v(t) = \begin{cases} 0 & \text{if } v \textit{ is inactive} \\ 1 & \text{if } v \textit{ is active} \end{cases} \tag{2.21}$$

Each edge in the IC model is weighted by an activation probability $P_{u,v}$, denoting the likelihood of node u activating node v. Additionally, one sets a threshold on the edges $\xi_{u,v}$ such that $\xi_{u,v} \in [0,1]$ to impede the independent influence of the nodes. Thus, node v is activated by node u if and only if:

$$P_{u,v} \geq \xi_{u,v} \tag{2.22}$$

One can consider the IC model as a variant of the SIR model with a varying infection rate among the nodes ($P_{u,v}$) instead of a constant infection rate (λ) between all the individuals. Additionally, one can consider that the IC model has a recovery rate $\psi = 1$ for all the nodes since each node has a single chance of activating its neighbor(s) [73]. Apart from marketing, It has been widely used in various situations, spanning from investigating rumor spreading [77], combating misinformation [78], sentiment analysis [79], and incorporation of textual data for better diffusion modeling [80].

2.3/ NETWORK QUALITY MEASURES

One can classify the several quality measures characterizing a community structure into two groups. Metrics in the first group exploit connectivity patterns [81, 82] while members of the second group rely on network models [83, 84]. This section briefly discusses these measures while emphasizing the overlapping modularity, which is used in Chapter 5.

The quality functions $f(c_q)$ are usually defined at the community level c_q. Then, the average on all the communities $f(C)$ characterizes the network's community structure. In the following, we recall the definitions for a community c_q.

2.3.1/ QUALITY MEASURES BASED ON CONNECTIVITY

One can divide quality measures based on connectivity into three categories based on the type of connectivity they consider: 1) measures based on internal connectivity, 2) measures based on external connectivity, and 3) measures based on internal & external connectivity. We briefly describe the main ones.

MEASURES BASED ON INTERNAL CONNECTIVITY

Measures based on internal connectivity quantify the quality of the community according to edges between nodes located inside the community.

1. Average internal degree is the average internal degree of the nodes included in c_q. The higher the value of the average degree, the stronger the community structure. It is defined as follows:

$$f(c_q) = \frac{2 \mid E^{in}_{c_q} \mid}{n_{c_q}}$$

2. Internal density measures the density of links joining nodes inside community c_q. The higher the value of the internal density, the stronger the community structure. It is defined as follows:

$$f(c_q) = \frac{2 \mid E^{in}_{c_q} \mid}{n_{c_q}(n_{c_q} - 1)}$$

MEASURES BASED ON EXTERNAL CONNECTIVITY

Measures based on external connectivity quantify the quality of the community based on links pointing out of the community.

1. Expansion quantifies the average proportion of external links per node inside c_q. The lower the value of expansion, the stronger the community structure. It is defined as follows:

$$f(c_q) = \frac{\mid E_{c_q}^{out} \mid}{n_{c_q}}$$

2. Cut ratio is the ratio of external links out of all possible links. The lower the value of the cut ratio, the stronger the community structure. It is defined as follows:

$$f(c_q) = \frac{\mid E_{c_q}^{out} \mid}{n_{c_q}(N - n_{c_q})}$$

MEASURES BASED ON INTERNAL & EXTERNAL CONNECTIVITY

Measures based on internal and external connectivity take information from the internal and external links simultaneously to quantify a community's quality.

1. Average-Out Degree Fraction (Average-ODF) is the average fraction of edges connecting nodes in community c_q to others. The lower the value of the Average-ODF, the stronger the community structure. It is defined as follows:

$$f(c_q) = \frac{1}{n_{c_q}} \sum_{i \in c_q} \frac{k_i^{out}}{k_i}$$

2. Fitness function measures the fraction of internal links to the total number of links of the community [85]. The higher the internal degree of the community, the stronger the community. It is defined as follows:

$$f(c_q) = \frac{2 \mid E_{c_q}^{in} \mid}{(2 \mid E_{c_q}^{in} \mid + \mid E_{c_q}^{out} \mid)^{\alpha}}$$

where $\alpha > 0$ is a resolution parameter that tunes the size of the communities.

2.3.2/ QUALITY MEASURES BASED ON MODELS

Some quality measures rely on a specific criterion based on a model to quantify the community structure strength. These measures defined for non-overlapping communities have natural extensions for networks with an overlapping community structure. Modularity and the Potts model are the most influential. Indeed, researchers have widely used these measures as an optimization criterion to identify communities in a network [84, 86–90]. We give a general presentation of modularity and its various versions and present the Potts model briefly.

OVERLAPPING MODULARITY

Initially introduced by Newman [86], it assumes that a node belongs to a single community. In numerous real-world scenarios, this is not the case. Therefore, Nepusz*et al.* [83] proposed incorporating the community membership strength vector $\Theta(i)$ of a node i

to relax this requirement. Therefore, each node i possesses a community membership strength defined as follows:

$$\Theta(i) = (\theta_{i,c_1}, \theta_{i,c_2}, ..., \theta_{i,c_q}, ..., \theta_{i,c_{|C|}}) \tag{2.23}$$

Under this setting, the overlapping modularity definition for fuzzy and crisp communities is given by:

$$f(c_q) = \frac{1}{2 \mid E \mid} \sum_{i,j} \left[A_{i,j} - \frac{k_i k_j}{2 \mid E \mid} \right] s(\Theta(i), \Theta(j)) \tag{2.24}$$

where $s(\Theta(i), \Theta(j))$ is a function of the vectors of community membership strength of nodes i and j denoted as $\Theta(i)$ and $\Theta(j)$, respectively.

Modularity can be equivalently defined in terms of the intra-community links and inter-community links of each community [91]. Thus, equation 2.24 can also be written as:

$$f(c_q) = \frac{\mid E^{in}_{c_q} \mid}{\mid E \mid} - \left(\frac{2 \mid E^{in}_{c_q} \mid + \mid E^{out}_{c_q} \mid}{2 \mid E \mid} \right)^2 \tag{2.25}$$

Note that here modularity is defined for one community. When computed on the whole network, we denote it as $Q'(G)$.

Fuzzy overlapping community structure

In a fuzzy overlapping community structure, the community membership strength of a node is a value between zero and one, and the sum of the strength values is normalized [92]. Consequently, $\forall i \in V, \forall c_q \in C \; 0 \leq \theta_{i,c_q} \leq 1$ and $\sum_{c_q \in C} \theta_{i,c_q} = 1$.

This study uses two fuzzy definitions. The first one, called *reciprocity membership*, is independent of the network topology. The second, *degree membership*, exploits the network topology. The literature proposes various alternatives. For more information, the reader can consult the following reference [92].

Reciprocity membership represents a node's reciprocal number of communities. If a node i belongs to o_i communities, the community membership strength for these communities is given by:

$$\theta^R_{i,c_q} = \frac{1}{o_i} \tag{2.26}$$

It is equal to zero for the other communities. Note that no information is used about the topology of the overlap in this definition except that different communities share the overlapping node.

Degree membership incorporates information about the node's attachment to the various communities. It measures the fraction of links of a node reaching a specific community. It is defined as follows:

$$\theta^{D}_{i,c_q} = \frac{k_{i,c_q}}{k_i} \tag{2.27}$$

The question now is how to integrate the measure $s(\Theta(i),\Theta(j))$ of equation 2.24. In this study, we opt for the average of the nodes' community membership strengths to be incorporated within the adjacency matrix ($A_{i,j}$) for reciprocity membership and degree membership. This approach is proposed by Zhang *et al.* [93]. Based on the variables of equation 2.25, the adjacency matrix is weighted according to the following fuzzy definitions:

$$\mid E^{in}_{c_q} \mid = \frac{1}{2} \sum_{i,j \in c_q} \frac{\theta_{i,c_q} + \theta_{j,c_q}}{2} A_{i,j} \tag{2.28}$$

$$\mid E^{out}_{c_q} \mid = \sum_{i \in c_q} \sum_{j \in C - c_q} \frac{\theta_{i,c_q} + (1 - \theta_{j,c_q})}{2} A_{i,j} \tag{2.29}$$

Crisp overlapping community structure

In crisp overlapping community structure, nodes either fully belong to communities or do not [92]. Consequently, the membership strength of a node to a community is either 0 or 1. Consequently, $\forall i \in V, \forall c_q \in C$, $\theta_{i,c_q} \in \{0,1\}$ and $0 \leq \sum_{c_q \in C} \theta_{i,c_q} \leq \mid C \mid$.

In this case, the sum of the community membership strengths is equal to $\mid C \mid$ if node i belongs to all the communities, and it is equal to zero if it does not belong to any community. This study uses one crisp definition independent of the network topology called *node similarity*. Note that a non-overlapping community structure is a particular crisp overlapping case where all the community membership strengths of nodes are equal to zero except for a unique community, where it is equal to one.

Node similarity measures the similarity of two nodes depending on their community membership strength $\Theta(i)$. The idea is to account for all the binary memberships of the nodes in the vector $\Theta(i)$ and then convert the similarity between any two nodes to a fuzzy index. One can use any similarity measure [83]. In the following, we adopt the widespread cosine similarity:

$$s(\Theta(i),\Theta(j)) = cos(\Theta(i),\Theta(j)) \tag{2.30}$$

Two nodes with similar community membership exhibit a high cosine similarity index. Note that node similarity does not use any information about the topology of the overlap.

The cosine similarity value of the two community membership strength vectors is directly incorporated into the adjacency matrix if two nodes i and j belong to the same community. Based on the variables of equation 2.25, the adjacency matrix is weighted using the following fuzzy expressions:

$$\mid E^{in}_{c_q} \mid = \sum_{i,j \in c_q} s(\Theta(i),\Theta(j)) A_{i,j} \tag{2.31}$$

$$| E_{c_q}^{out} | = \sum_{i \in c_q} \sum_{j \in C - c_q} [1 - s(\Theta(i), \Theta(j))] A_{i,j} \tag{2.32}$$

POTTS MODEL

In the Potts model, an energy expression quantifies the quality of the communities considered as spin states [84]. Minimizing the system's energy expressed by the Hamiltonian of the spin model results in a well-separated community structure. It is defined as follows:

$$f(c_q) = - \sum_{i,j \in c_q} J(A_{i,j} - \gamma p_{i,j})$$

where:

- J is a multiplicative constant characterizing the weights between the nodes
- γ weights the importance of the energy from connected and missing edges
- $p_{i,j}$ can be any arbitrary null model denoting the probability of node i connecting to node j

MAP EQUATION

The map equation originating from information theory aims to find cohesive structures networks based on the flow of information [94, 95]. The underlying mechanism is based on compressing regions where a random walker has a higher probability of staying in for an extended period. By minimizing the map equation over the possible partitions in the network, communities can be revealed depending on how information flows. The map equation is defined as follows:

$$f(c_q) = q_{\curvearrowright} H(\mathcal{Q}) + p_{\circlearrowleft}^{c_q} H(\mathcal{P}^{c_q})$$

where:

- $q_{\curvearrowright} H(\mathcal{Q})$ denotes the average number of bits describing the movement between community c_q and the remaining communities
- $p_{\circlearrowleft}^{c_q} H(\mathcal{P}^{c_q})$ denotes the average number of bits describing the movement within community c_q

2.4/ SUMMARY

This chapter introduces the concepts and elements essential to understanding the thesis. It first discusses the centrality measures the thesis investigates, including classical and community-aware measures. The former is divided into neighborhood-based,

path-based, and iterative refinement-based measures, while the latter is divided into non-overlapping and overlapping measures. The chapter then presents four diffusion models, two belonging to epidemic models and two to information diffusion models. Lastly, the chapter discusses various network quality measures that can be used to describe networks, focusing on overlapping modularity, which forms the basis of the framework proposed in Chapter 5.

3

Diffusion on Networks with Community Structure

Contents

3.1/ Introduction

Diffusion begins from nodes located in specific areas of the network and spreads out with time. How to select the seed nodes to maximize diffusion is a fundamental problem. Centrality measures are one of the main approaches to do so. They rely on topological information from the network to quantify node importance. Since the community structure impacts the diffusion spreading dynamics [15, 17, 96], researchers showed that classical centrality measures may fallback in terms of selecting the most influential nodes [20–27]. Therefore, it is important to incorporate community structure information to select seed nodes that maximize diffusion.

Unlike the classical centrality measures, which focus more on either the local or the global influence of a node, the so-called community-aware centrality measures incorporate the node's local and global influence through its intra-community and inter-community links, respectively [20–27]. The difference between these measures is how they combine intra-community links and inter-community links, as shown in Chapter 2. If more importance is given to the intra-community links (i.e., local influence), the measure emphasizes hub-like nodes. On the other hand, if importance is given to the inter-community links (i.e., global influence), the measure renders more important bridge-like nodes.

The diffusive ability of the community-aware centrality measures in selecting seed nodes is assessed in a dynamic spreading scenario with specific conditions set on nodes and/or edges. Most of the studies use the SIR model to assess the impact of the selection of seed nodes either to maximize diffusion or to minimize it (this can also be called immunization) [20–26]. Despite being widely used, the SIR model does not convey all

real-world spreading scenarios. In particular, in the SIR model, a node can infect its neighborhood several times. In other words, a node has many chances to infect or influence its neighbor(s) before it is removed from the network. Nevertheless, sometimes the diffusion of a disease or a piece of information to raise awareness can be spread by a node once. That is to say, a node has a single chance to influence its neighbor(s). For instance, consider people meeting in a manifestation. They will meet in this manifestation once, and they may not meet again afterward. The piece of information from one person to another will be transmitted given this one-time chance. Another example is that a person may change his/her opinion towards a cause only if a sufficient number his/her neighbors adopts this opinion. The presence of various conditions that can occur in nodes or edges in the real world necessitates the creation of multiple diffusion models.

This grants us permission to pose the main research question of this chapter: how does the diffusion models' output depend on the seeds and the network? The seeds are selected based on the community-aware centrality measures. The community-aware centrality measures rely on the network's structure. The model's output (i.e., the diffusion spread) depends on the network's structure and the seed nodes. Thus, we investigate the interplay between the spread of various diffusion models, initiated through the seed nodes selected by the community-aware centrality measures, and the network's structure. This problem is relevant to many disciplines, from biology and epidemics to sociology and economics. In addition to the diffusion models, there are insufficient studies using a multiple-spreading phenomenon under a spreading scenario rather than in immunization. Another issue is that the community structure changes using one community detection algorithm over the other, which may impact the diffusion dynamics. This also poses a challenge on how the spread will evolve, given that the same seed nodes initiate the diffusion. Finally, we do not have a clear idea of how the community-aware centrality measures compare in controlled synthetic networks and diverse real-world networks. Indeed, previous works mainly focused on a small set of synthetic and real-world networks. This does not enable us to rigorously answer when community-aware centrality measures outperform and what their bottlenecks are.

Thus, despite their outperformance, the literature so far renders a blurred vision of community-aware centrality measures, their robustness, and their limitations. All of the stated challenges are tackled in this chapter. To conduct the study, we systematically use eight community-aware centrality measures on a set of four conceptually different diffusion models using a set of synthetic and real-world networks from diverse domains under the multiple-spreader scheme. Therefore, three main parameters are under investigation: the diffusion model and its underlying dynamics, the network and its characteristics, and the community-aware centrality measures and their inner workings. The employed diffusion models are the Susceptible-Infected (SI) model, Susceptible-Infected-Recovered (SIR) model, the Linear Threshold (LT) model, and the Independent Cascade (IC) model. Synthetic networks are generated using the Lancichinetti, Fortunato, and Radicchi algorithm (LFR) [97] where several parameters can be varied, including the community structure strength, the community size distribution, and the degree distribution. In real-world networks, Infomap and Louvain community detection algorithms are used to uncover the underlying community structure, and their impact is also studied. The community-aware centrality measures, the diffusion models, and the networks are discussed in more detail in Chapter 2.

In this chapter, we contribute to the literature by:

1. Evaluating the community-aware centrality measures selecting seed nodes in four distinct diffusion models.
2. Using a set of 45 LFR synthetic networks to investigate the impact of the community structure strength, community size distribution, and degree distribution.
3. Applying the methods on a large set of real-world networks for an extensive statistical study.
4. Investigating two fundamentally different community detection algorithms to investigate their impact on dynamics.

3.2/ STATE OF THE ART

Influential nodes are critical in boosting or curbing spreading phenomena in complex networks. A multitude of classical centrality measures has been proposed to quantify node influence. These measures prove their merit in many scenarios, like assessing the infectious capacities of nodes [14] to quantifying financial distress [98] and applying viral marketing [12]. Researchers have shown that classical centrality measures may undermine the influence of nodes in networks with community structure [20–27]. Indeed, many real-world networks are characterized by a community structure that drastically impacts spreading dynamics [15, 17, 96]. Thus, in networks with a community structure, nodes that may not be considered influential by a classical centrality measure (i.e., agnostic about the community structure) may be of ultimate influence when one considers the mesoscopic organization of the network.

The exploitation of communities to identify influential nodes using centrality dates back to 2005 when Guimerà and Amaral [20] proposed the Participation Coefficient, which uncovered the key metabolites across species in metabolic networks. Zhao *et al*. [21] proposed the Community-based Centrality, capable of identifying influential nodes in which the classical degree, betweenness, and eigenvector centralities could not identify in the Susceptible-Infected-Recovered (SIR) model with a single-spreader scheme. Unlike Community-based Centrality, Comm centrality proposed by Gupta *et al*. [22] adapts to the network's strength of community structure, succeeding in identifying hubs and bridges, with the latter being prioritized in an immunization scenario using SIR. Luo *et al*. [23] merged the network's community structure and hierarchy to develop K-shell with Community, proving its outperformance against classical centrality measures in the SIR model with a single-spreader scheme. Tulu *et al*. [24] showed that using the entropy of a node's intra-community and inter-community links, nodes disseminating information quickly can be better identified in the SIR model. Ghalmane *et al*. [25] proposed Community Hub-Bridge, which showed its effectiveness in hindering an epidemic by immunizing influential nodes under the SIR dynamics in networks with a strong community structure. Magelinski *et al*. [26] exploited the so-called modularity, a quality measure to assess the community structure of a network to identify hubs and bridges. The authors showed that their community-aware centrality could dismantle a very large infrastructural network eight times more effectively than other centrality measures by taking a limiting case of the SIR model. Recently, Blöcker *et al*. [27] showed the merit of an information-theoretic community-aware centrality measure based on the map equation in the SIR model using a single-spreader scheme and the Linear Threshold (LT) model using a multiple-spreader scheme.

Despite the outperformance of the community-aware centrality measures compared to classical ones in identifying influential nodes, several limitations need to be addressed. First, most of the community-aware centrality measures are assessed exclusively under the SIR dynamics, either to maximize diffusion [21, 23, 24, 27] or minimize it [22, 23, 25, 26] by removing the most central nodes. The latter case is also referred to as "immunization." Nevertheless, the SIR model does not characterize all situations. Despite researchers' aim to develop generalized diffusion models, many cases entail adding specific conditions that are not applicable in all real-world situations. Therefore, many diffusive models exist to characterize better cases occurring in the real world. Only one community-aware centrality measure, the Map Equation Centrality [27], is assessed using the SIR and LT dynamics. Second, most of the studies maximizing SIR diffusion use the single-spreader scheme. To minimize diffusion, studies naturally select multiple nodes [22, 23, 25, 26]. That being said, Participation Coefficient [20] is the only community-aware centrality measure not assessed with respect to a dynamic model since its original aim was to identify key proteins and construct cartography of metabolic networks rather than compare the measure's diffusive power. Third, many studies use a single community detection algorithm if the network's community structure is unknown. Therefore, it is not well understood how the mesoscopic arrangement of communities affects the dynamics within the same network. Finally, every community-aware centrality measure is assessed on a small sample of real-world and synthetic networks. Table 3.1 summarizes the works concerning the development and comparison of community-aware centrality measures. The limitations raise several concerns regarding the consistency of the community-aware centrality measures, and this chapter aims to address these questions.

Community-aware centrality measures	Diffusion model & goal	Node selection method	Real-world networks	Synthetic networks	Number of community detection algorithms
Participation Coefficient [20]	-	-	12	-	1
Community-based Centrality [21]	SIR ↗	Single	6	-	5
Comm Centrality [22]	SIR ↘	Multiple	4	3	1
K-shell with Community [23]	SIR ↗ ↘	Single & Multiple	4	-	1
Community-based Mediator [24]	SIR ↗	Single & Multiple	5	2	1
Community Hub-Bridge [25]	SIR ↘	Multiple	6	5	3
Modularity Vitality [26]	SIR ↘	Multiple	2	3	1
Map Equation Centrality [27]	SIR ↗ & LT ↗	Single & Multiple	12	1	2

Table 3.1: A summary of the studies of community-aware centrality measures. SIR means Susceptible-Infected-Recovered model and LT refers to Linear Threshold model. The character '-' refers to "not applicable". ↘ indicates that the goal is to minimize diffusion and ↗ indicates that the goal is to maximize diffusion.

3.3/ SYNTHETIC NETWORKS

This section investigates the interplay between the network's community structure, various diffusion mechanisms based on the models provided in Chapter 2, and community-aware centrality measures on a set of synthetic networks generated by the Lancichinetti, Fortunato, and Radicchi (LFR) algorithm [97]. This algorithm allows the tuning of various parameters of the community structure. We investigate the influence of the community structure strength controlled by the mixing parameter (μ), the degree distribution power-law exponent (γ), and the community size distribution power-law exponent (θ). More details about the synthetic networks and the parameters set to generate them are provided in Appendix I.

INFLUENCE OF THE COMMUNITY STRUCTURE STRENGTH

The mixing parameter (μ) controls the community structure strength. Low values yield networks with a strong community structure since few inter-community links exist. As μ increases, the network's community structure strength loosens, resulting in the disappearance of dense and well-defined regions. To study the effect of the community structure strength on the various diffusion dynamics, we generate LFR networks with strong (μ = 0.05), medium (μ = 0.20), and weak (μ = 0.70) community structures. Given the ranking of a community-aware centrality measure, a fraction f_o of the top-ranked nodes in the network is initially infected/activated in each of the SI, SIR, IC, and LT models. The results are reported in Fig 3.1. The evaluation measure in the SI model is the average number of iterations needed for a given f_o to infect 50% of the network. The lower the number of iterations, the more effective the centrality measure. In the SIR and IC models, the relative outbreak/activation size (i.e., $\Delta R/\Delta A$) is computed. This value quantifies the difference between the number of nodes recovered/activated at the end of the dynamical process when f_o is based on a given community-aware centrality measure and a baseline measure which is the degree centrality. Recall that $\Delta R = \frac{R_c - R_r}{R_r}$ (see Appendix I for more details). The higher it is, the better the performance of the community-aware centrality measure. Finally, in the LT model, the evaluation measure is the total number of activated nodes normalized by the size of the network (i.e., the activation rate Ar).

Two main phenomena dominate as the community structure strength (μ) varies from strong (μ = 0.05) to weak (μ = 0.70). First, the stronger the community structure, the more pronounced the difference is in the performance of the community-aware centrality measures. As the community structure strength decreases (i.e., weakens), the performance of the community-aware centrality measures becomes more comparable, and differences are less visible. Community-aware centrality measures are well-adapted to networks with a well-defined community structure. With this structure, each measure can exploit various community information to identify influential hubs and bridges that contribute to the network's community structure. If the community structure is loosely defined, it becomes more difficult for the community-aware centrality measures to pinpoint these influential nodes. Indeed, in a weak community structure, hubs and bridges become less prominent, and the average degree of the nodes becomes more analogous.

The second phenomenon is related to the divergence in the scales' magnitude while the dynamical processes take place. In the SI model, with a strong community structure, the epidemic diffusion needs more iterations to reach 50% of the network. As communities

share a few inter-community links in a network with a strong community structure, the infection tends to stay more localized in the communities. With a decrease in the community structure strength, the proportion of inter-community links increases. Therefore, the infection can spread more quickly to the remaining communities. Thus, fewer iterations are needed to infect half of the network. In the SIR and IC model, it is clear that when the network has a strong community structure, a set of community-aware centrality measures outperform degree centrality by a large difference. However, as the community structure strength weakens, the community-aware centrality measures become more comparable to the performance of the degree centrality.

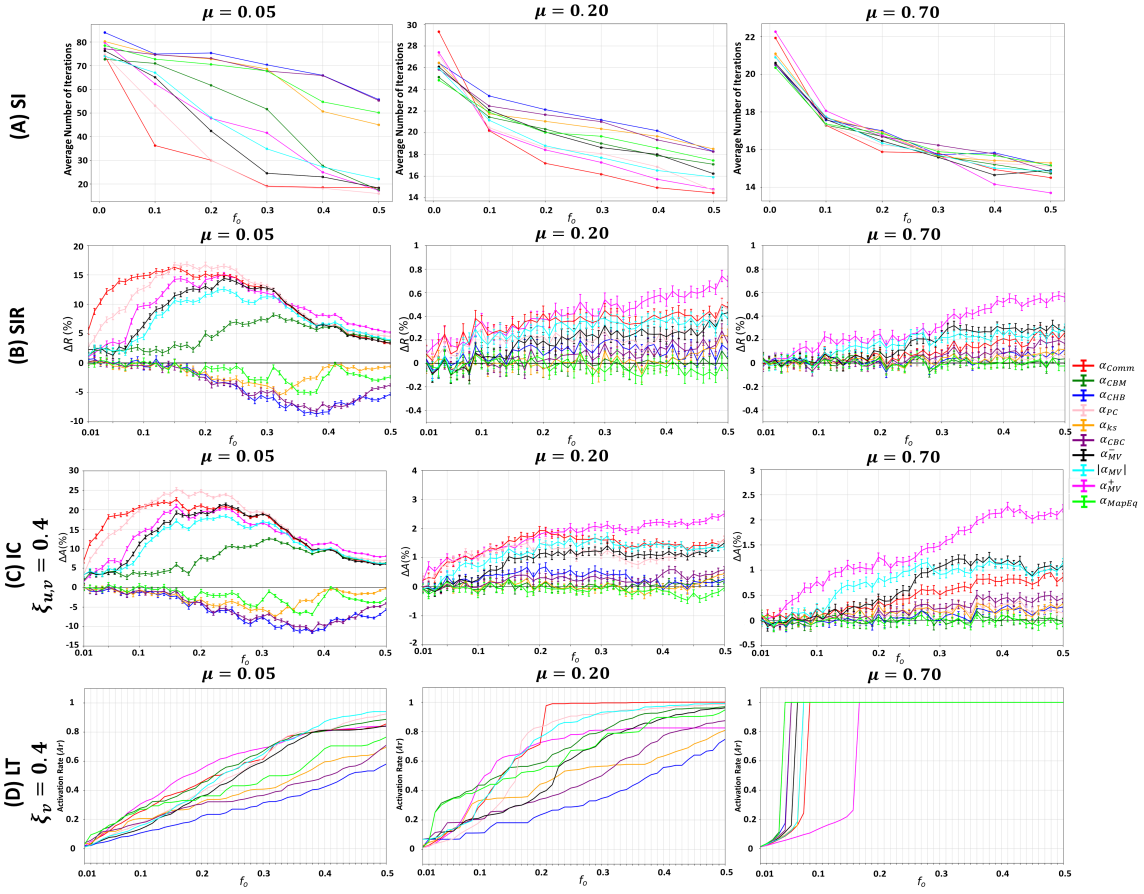


Figure 3.1: **Behavior of the community-aware centrality measures under various dynamic models in synthetic networks while varying the mixing parameter (μ).** The first, second, third, and fourth rows indicate the results of the (A) SI model, (B) SIR model, (C) IC model, and (D) LT model.

Inspecting the community-aware centrality measures in more detail, most of these measures are a variant of degree centrality exploiting the inter-community and intra-community links in various ways. The smaller the difference between these two types of links - usually prevailing in a network with a loose community structure - the higher the resemblance of the community-aware centrality measures to degree centrality. Therefore, in a network with a weak community structure, the outperformance of the community-aware centrality measures is insignificant compared to degree centrality. The latter is advised for usage as it does not need community-level information. Nevertheless, with a

strong community structure strength, community-aware centrality measures can extract information that the community-agnostic degree centrality cannot.

An important finding can be extracted independently of the community structure strength (μ). It concerns the similarities of the various dynamic models. By visually inspecting Fig 3.1, one can note that generally, at any given f_o, the top 2 outperforming centrality measures are comparable across the SI, SIR, and IC models, excluding the LT model. For instance, at f_o = 0.05 and μ = 0.05, the top 2 outperforming centrality measures are Comm Centrality (α_{Comm}) and Participation Coefficient (α_{PC}) in the SI, SIR, and IC models. In contrast, Map Equation Centrality (α_{MapEq}) and Community-based Mediator (α_{CBM}) are the top 2 most performing in the LT model at f_o = 0.05 and μ = 0.05. This behavior is logical as the SI, SIR, and IC are a variant of one another. The SIR is the SI with an additional "recovered" state. The IC sets thresholds on edges, and nodes have one chance to infect/activate their neighbors, while in the SIR model, a node has more than one chance. Even though differences exist, they are nominal. Indeed, their dynamics follow the simple contagion dynamics where nodes getting activated/infected are independent of their surroundings. This is not true for the LT model, where a node's activation depends on its neighborhood's aggregate activations. Subsequently, activations are harder to diffuse across the network, especially if the network has a strong community structure strength [99, 100].

These results suggest that the community-aware centrality measures are more profitable in networks with a strong community structure strength. They also suggest that one should be prudent in using the measures even with a strong community structure strength, as the outperformance depends on the model. Some measures are well-suited to the SI, SIR, and IC models, while others are more suited to the LT dynamics.

Influence of the community size distribution exponent

The community size distribution exponent (θ) is responsible for the frequency and the size of the generated communities. We fix the community structure strength at μ = 0.05 and generate three networks with three different community size distribution exponents. The first, having θ = 2, indicates that large communities make up most of the network, with the existence of few small communities, resulting in a large variance in the community sizes. The second, having θ = 2.7, yields less variance in the community sizes with a larger number of communities. Finally, the third, having θ = 3, a high number of communities exist with equivalent sizes. Given the ranking of a community-aware centrality measure, a fraction f_o of the top-ranked nodes in the network is initially infected/activated in each of the SI, SIR, IC, and LT models. The results are reported in Fig 3.2.

In case the dynamics follow the SI, SIR, or IC, it can be noticed that the general trends of the community-aware centrality measures persist whether the network is generated with θ = 2, θ = 2.7, and θ = 3. The main difference is in the magnitude of the final output of each of the models. However, with the LT model, the behavior of the community-aware centrality measures changes with every θ under investigation.

Particularly, in the SI model, in the network with a larger variance in the community size distribution (i.e., at θ = 2), it takes less time to infect 50% of the network compared to networks with fewer communities of equivalent sizes. For instance, when θ = 2, at f_o = 0.2, the average number of iterations for Comm Centrality (α_{Comm}), the best performing centrality, takes 20 iterations while with θ = 2.7, it takes 30 iterations and with θ = 3 it

takes 32 iterations. The magnitude of the relative outbreak size (ΔR) in the SIR model shows that the outperformance of Comm Centrality (α_{Comm}), Modularity Vitality targeting hubs (α^{+}_{MV}), and Participation Coefficient (α_{PC}) is more pronounced in the networks with θ = 2.7 and θ = 3. For instance, let's take f_o = 0.10, ΔR of Comm Centrality amounts to 8.5% in the network with θ = 2, while in the networks with θ = 2.7 and θ = 3, ΔR amounts to 15%. Under the IC dynamics, the highest magnitude of the relative activation size (ΔA) is reached when θ = 2 (ΔA = 13%). In contrast, when θ = 2.7 and θ = 3, the maximum ΔA reached amounts to 25% and 28.5%, respectively.

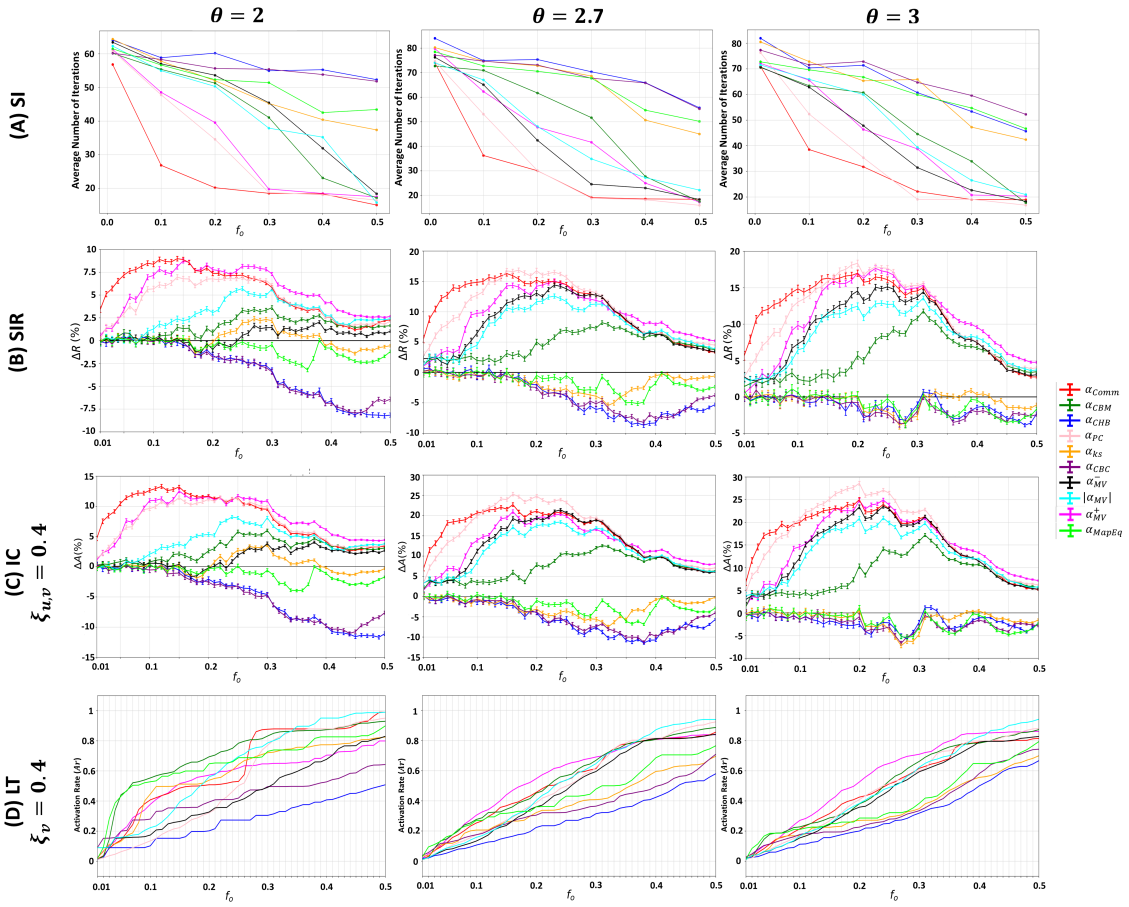


Figure 3.2: **Behavior of the community-aware centrality measures under various dynamic models in synthetic networks while varying the community size distribution exponent (θ).** The first, second, third, and fourth rows indicate the results of the (A) SI model, (B) SIR model, (C) IC model, and (D) LT model.

Despite the similarities in the general trends of the community-aware centrality measures, one subtle difference needs to be noted concerning Participation Coefficient. In the three models, Participation Coefficient performs less when θ = 2. As stated earlier, θ = 2 indicates that the network is characterized by a few small communities and many large communities, which comprise most of the network. Participation Coefficient's effectiveness depends on the number of communities within the network. If there are many communities, it has more room to distinguish the difference in influence between nodes. On the other hand, having fewer communities makes it less effective as many nodes will have similar centrality, making it difficult to distinguish their influence characteristics.

In the LT model, the impact of adjusting θ is more pronounced on the performance of the top-performing community-aware centrality measures in terms of behavior rather than magnitude. The contrast becomes more noticeable when the budget range goes from low to medium. When the budget is high, in all of the studied θ, the strategy is to target hubs and bridges together using Modularity Vitality targeting hubs and bridges ($|\alpha_{MV}|$). We now discuss when the budget spans from low to medium. When the network has a significant variance in the community size distribution ($\theta = 2$), hub-like nodes are preferred up to a small value for the budget availability, then bridge-like nodes are preferred. In contrast, with a smaller variance in the community size distribution ($\theta = 2.7$ and $\theta = 3$), hub-like nodes are always preferred. More specifically, when the value of θ is equal to 2, the hub-like nodes that produce the highest outbreak until f_o reaches 0.05 are selected by Map Equation Centrality (α_{MapEq}) and Community-based Centrality (α_{CBC}). From f_o = 0.06 to f_o = 0.34, bridge-like nodes selected by Community-based Mediator (α_{CBM}) and then by Comm Centrality (α_{Comm}) are the nodes that generate the highest activation rate. If θ is equal to 2 or 2.7, the hub-like nodes preferred at the small budget are chosen by Map Equation Centrality (α_{MapEq}), and then Modularity Vitality targeting hubs (α^{+}_{MV}) takes over for a broader range of budget availabilities.

To sum up, the results of the SI, SIR, and IC models suggest that changing the community size distribution exponent has a greater community-aware centrality measures' magnitude in the model's output rather than their behavior. In networks with a large variance in community sizes, the outbreak size in the SIR model and the activation size in the IC model are not as pronounced as in networks with a smaller variance, implying that the outbreak can more easily spread to many communities with equivalent sizes. However, many communities may remain unaffected if the infection starts in big communities and remains within them. In the SI model, when there is a large variance in community sizes, it takes less time to infect 50% of the network since it consists of only a few big communities. Suppose many nodes are in the same community, making up almost 50% of the network. In that case, it is easy to infect/activate that community because the nodes in the community are more likely to be directly or indirectly connected. In the LT model, the community size distribution has a greater effect on the behavior of community-aware centrality measures than their magnitude. When there is a large variance in community size distribution, bridge-like nodes play a crucial role in the medium budget range, as the selected bridges are likely to be located in large communities, resulting in a higher outbreak. Conversely, when there are many communities of similar sizes, it is more beneficial to target hub-like nodes since the influence of bridge nodes may stop at the border of a community with high density. These findings are supported by studies on contagion and networks [99, 100].

Influence of the degree distribution exponent

The degree distribution exponent (γ) characterizes the number of links nodes have in a network. The LFR algorithm generates networks with a power-law degree distribution fitting the degree distribution of many real-world networks [101, 102]. Many real-world networks are distinguished by γ falling between 2 and 3 [103, 104]. To investigate three representative cases, we fix the community structure strength at $\mu = 0.05$ and generate three networks with γ = [2, 2.7, 3]. While preserving the community structure, the network portrays a hub-and-spoke structure when $\gamma = 2$ [105]. On the other extreme, the nodes inside the communities have more comparable degrees, resembling a random-like network

when γ = 3. At γ = 2.7, the network resembles a typical scale-free network. Given the ranking of a community-aware centrality measure, a fraction f_o of the top-ranked nodes in the network is initially infected/activated in each of the SI, SIR, IC, and LT models. The results are reported in Fig 3.3.

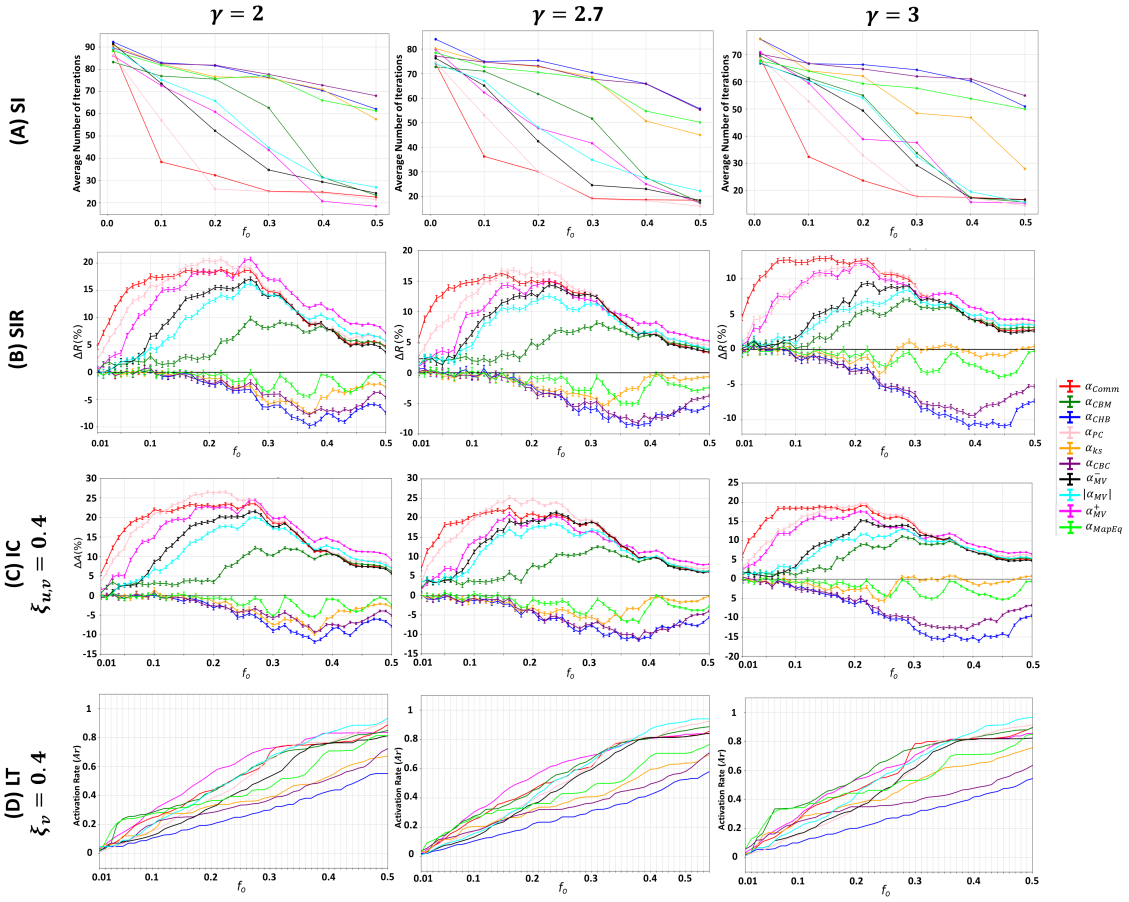


Figure 3.3: **Behavior of the community-aware centrality measures under various dynamic models in synthetic networks while varying the degree distribution exponent (γ).** The first, second, third, and fourth rows indicate the results of the (A) SI model, (B) SIR model, (C) IC model, and (D) LT model.

Similar to the variation of the community size distribution exponent (θ), the general trend persists when varying the degree distribution exponent (γ) in the SI, SIR, and IC models where the difference is attributed to the magnitude of the models' output, while with the LT model, the behavior of the community-aware centrality measures is what changes rather than the magnitude.

To begin with the SI model, the time it takes to infect 50% of the network decreases as the initial fraction of infected nodes (f_o) increases, as γ increases. Indeed, the network structure impacts the number of iterations it takes to infect 50% of the network. The random-like structure inside the communities, as found in γ = 3 networks, results in a faster spread compared to the hub-and-spoke-like structure, as found in γ = 2 networks and to a lesser extent when γ = 2.7. For example, when f_o = 0.1, the best performing centrality in the SI model, Comm Centrality (α_{Comm}), takes an average of 39 iterations to infect 50% of the network in a γ = 2 network, 36 iterations in a γ = 2.7 network, and 31.5

iterations in a γ = 3 network. In the SIR and IC models, the output of all the measures performs more in the networks with θ = 2 and θ = 2.7 compared to γ = 3. Let's take f_o = 0.10 following the SIR model, ΔR of Comm Centrality amounts to 17% in the network with γ = 2, to 16% in the network with γ = 2.7, and to 13% in the network with γ = 3. Under the IC dynamics, the maximum relative activation size (ΔA) when it equals 2 is 26.5%, while at γ = 2.7 and γ = 3, the ΔA reached is 25% and 19.5%, respectively.

Although there are similarities in the overall patterns of the community-aware centrality measures in the SI, SIR, and LT models, the Participation Coefficient is influenced by changes in the degree distribution exponent, similar to the impact of changes in the community size distribution exponent. Specifically, the Participation Coefficient performs better when the network is created with γ = 2 and γ = 2.7. This suggests that the Participation Coefficient benefits from having differences in node degrees, which allows it to distinguish between nodes and identify the most influential ones.

The results of the LT model show that the main difference between different values of γ is observed when the budget availability is medium. When the budget is high (i.e., $f_o \geq$ 0.40), targeting hub-like and bridge-like nodes using Modularity Vitality targeting hubs and bridges ($|\alpha_{MV}|$) is always the most effective strategy, regardless of γ. Similarly, when the budget is low (i.e., $f_o \leq$ 0.05), it is always better to target hub-like nodes selected by Map Equation Centrality (α_{MapEq}). However, when the budget is medium, networks with γ = 2 and γ = 2.7 tend to benefit more from targeting hub-like nodes using Modularity Vitality targeting hubs (α^{+}_{MV}). On the other hand, in networks with γ = 3, where the communities are more random, bridge-like nodes become more influential. Community-based Mediator (α_{CBM}) selects nodes that are well-connected between different communities in the network for a higher activation rate in this case.

In brief, results show that community-aware centrality measures exhibit consistent behavior across the SI, SIR, and IC models as the degree distribution exponent changes. However, the models' output based on these measures varies in magnitude. When the degrees of nodes in communities are similar, the SI model takes less time to infect more of the network. But in the SIR and IC models, community-aware centrality measures are comparable to degree centrality as they are evaluated using the relative outbreak and activation sizes. This is because the measures have less power to differentiate between hub-like and bridge-like nodes when node degrees are similar. In contrast, the LT model's community-aware centrality measures exhibited differences in behavior rather than magnitude, particularly at medium budget availabilities. Results indicate that targeting bridge-like nodes is better when node degrees are comparable in their communities. This is because similar degrees may imply similar influence, making differentiation difficult. Therefore, selecting bridge-like nodes has a better chance of igniting a more significant impact in the network.

3.4/ REAL-WORLD NETWORKS

In this section, we investigate the interplay between the diffusion dynamics of the models provided in Chapter 2 and the community-aware centrality measures on real-world networks. Unlike synthetic networks, the topological characteristics of real-world networks cannot be controlled. Indeed, real-world networks are characterized by diverse structures that affect the diffusion dynamics differently. Moreover, these networks pertain to various

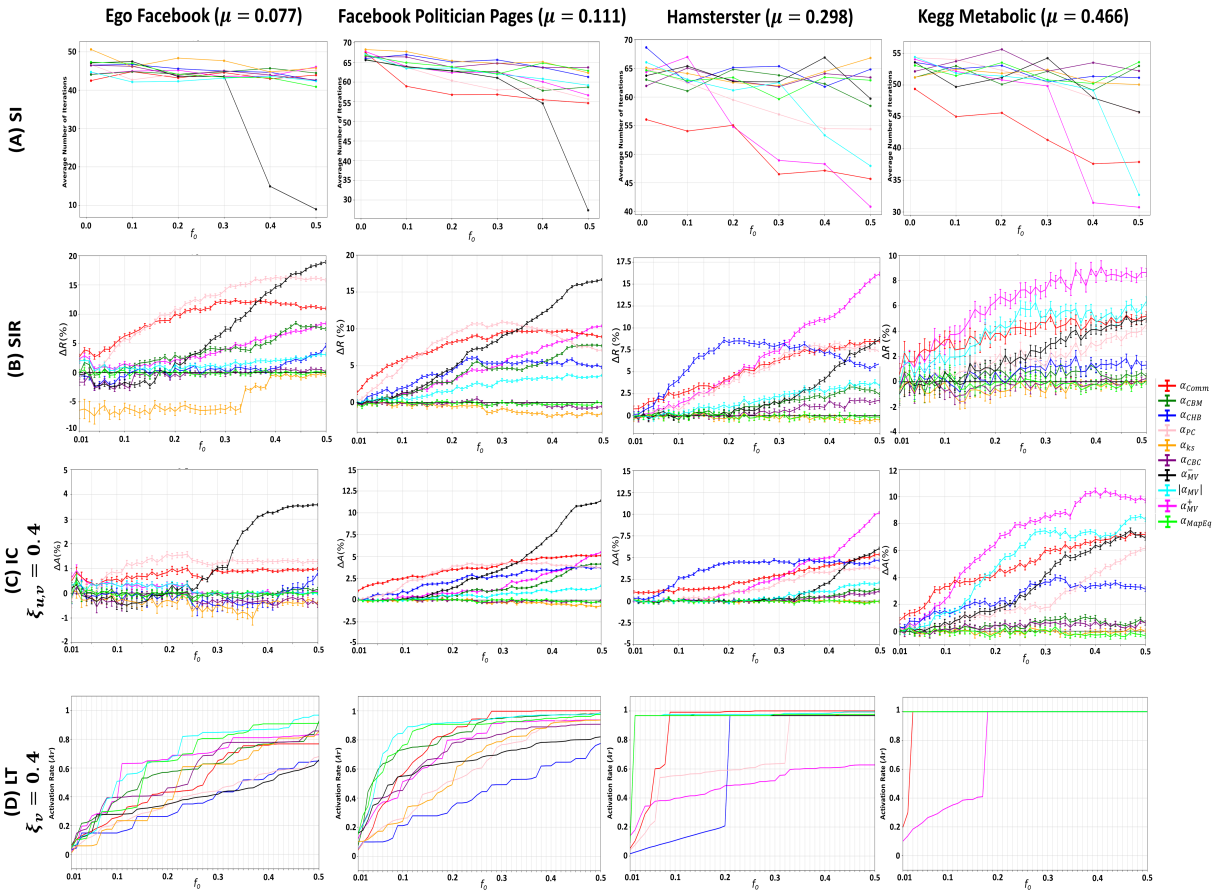


Figure 3.4: **Behavior of the community-aware centrality measures under various dynamic models in real-world networks with varying community structure strengths.** The first, second, third, and fourth rows indicate the results of the (A) SI model, (B) SIR model, (C) IC model, and (D) LT model.

domains (i.e., infrastructural, social, acting, biological, and collaborative). Thus, nodes and edges have specific roles in maintaining the normal functioning of the network. Since the community structure of real-world networks is unknown a priori, we uncover their communities using the Infomap [106] community detection algorithm. At a later stage, we investigate the consistency of the results using the Louvain [87] community detection algorithm.

Similar to the methodology adopted with synthetic networks, a fraction f_o of the top-ranked nodes in each network is initially infected/activated, given the ranking of a community-aware centrality measure in each of the SI, SIR, IC, and LT models. For brevity, the results of four real-world networks are reported in Fig 3.4. These networks are representative cases of the dynamics seen across the networks under study. Indeed, the extensive analysis across the models shows two network categories in every diffusion model investigated. These two categories, illustrating a different behavior in terms of the spreading dynamics, can be divided based on the network's community structure strength.

The first category comprises networks with medium to weak community structure strengths, such as the networks Hamsterster (μ = 0.298) and Kegg Metabolic (μ = 0.466). In this category, a shared trend illustrates that up to a certain fraction of initially

infected/activated nodes (f_o), bridge-like nodes using Comm Centrality (α_{Comm}) outperform the remaining measures. After passing f_o, which is network-dependent, hub-like nodes using Modularity Vitality targeting hubs (α^{+}_{MV}) outperform other measures in terms of spreading capability in each of the SI, SIR, and IC models. In between, Community Hub-Bridge (α_{CHB}) provides good results in a medium range of f_o only in the SIR and IC models. Results in the LT model diverge from the remaining models. With the LT dynamics, hub-like nodes using Map Equation Centrality (α_{MapEq}) outperform the remaining measures up to a certain f_o. Then, it either persists in its outperformance with other measures or Comm Centrality (α_{Comm}), which has a preference for bridge-like nodes, outperforms the remaining measures (as seen with Hamsterster).

The second category comprises networks with a strong community structure strength, such as Ego Facebook (μ = 0.077) and Facebook Politician Pages (μ = 0.111). Within this category, under the SI, SIR, and IC dynamics, bridge-like nodes always yield the highest performance. The distinction lies in which community-aware centrality measure yields such performance. Generally, Comm Centrality (α_{Comm}) has the highest performance up to a certain f_o. Then, Participation Coefficient (α_{PC}) overcomes Comm Centrality (α_{Comm}) only in the SI and SIR dynamics. Then, in the SI, SIR, and IC dynamics, Modularity Vitality targeting bridges (α^{-}_{MV}) outperforms all remaining measures. The LT dynamics pose different outcomes. At first, hub-like nodes using Community-based Centrality (α_{CBC}) or Map Equation Centrality (α_{MapEq}) outperform the remaining measures. After exceeding a certain f_o, several measures may show high performance, namely Community-based Mediator (α_{CBM}), Modularity Vitality targeting hubs (α^{+}_{MV}), Modularity Vitality targeting hubs and bridges ($|\alpha_{MV}|$), Map Equation Centrality (α_{MapEq}), and/or Comm Centrality (α_{Comm}).

A divergence in scale occurs among the real-world networks. For instance, the performance of the community-aware centrality measures in networks with a weak community structure is more comparable with degree centrality than in networks with a strong community structure. However, the divergence in scale is less significant than in synthetic networks. Moreover, unlike synthetic networks, differences in the performance of the community-aware centrality measures exist in real-world networks under the SI, SIR, and IC dynamics. A high variance between the curves is seen regardless of whether the network has a strong or weak community structure strength.

The study's results show that the SI, SIR, and IC models behave similarly to synthetic networks, but the dynamics differ in real-world networks based on the strength of their community structure. In networks with a weak community structure, bridge-like nodes lead to higher outbreaks/activations until hub-like nodes perform better. However, bridge-like nodes always perform better in networks with a strong community structure. Regarding the LT dynamics, hub-like nodes outperform other measures up to a certain f_o, and after that, other measures with preferences for hub-like, bridge-like, or both types of nodes show better performance. Additionally, real-world networks have a more pronounced variance between the curves regardless of their community structure strength, showcasing their unique characteristics that diversely affect the network's dynamics. This contrasts with synthetic networks, where all parameters are controlled.

3.5/ DISCUSSION

In this section, we address questions related to why the results of the dynamic models seen with the real-world networks are obtained. We refer to the fraction of initially infected/activated nodes (f_o) as "budget availability" thenceforth.

(1) Why is it more beneficial to target bridge-like nodes at low budget availability and hub-like nodes at high budget availability in the SI, SIR, and IC diffusion models when the network has a medium to weak community structure strength? As it was previously seen, in networks with a medium to weak community structure strength, Comm Centrality (α_{Comm}) generally results in the highest outbreak when the budget is limited. To investigate why we take the Kegg Metabolic network, which has a weak community structure strength (μ = 0.466). Using this network, we compare in Fig 3.5 the position of the top nodes chosen based on low (i.e., f_o = 1%), medium (i.e., f_o = 25%), and high (i.e., f_o = 40%) budget availabilities. For comparison purposes, we take the various budget availabilities according to the rankings based on Comm Centrality (α_{Comm}), K-shell with Community (α_{ks}), and Modularity Vitality targeting hubs (α^{+}_{MV}). As we can see in Fig 3.5, Comm Centrality (α_{Comm}) targets nodes distributed across the network when the budget is either low (i.e., f_o = 1%) or medium (i.e., f_o = 25%). These nodes yield a higher spreading capability in the SI, SIR, and IC models, as Fig 3.4 shows.

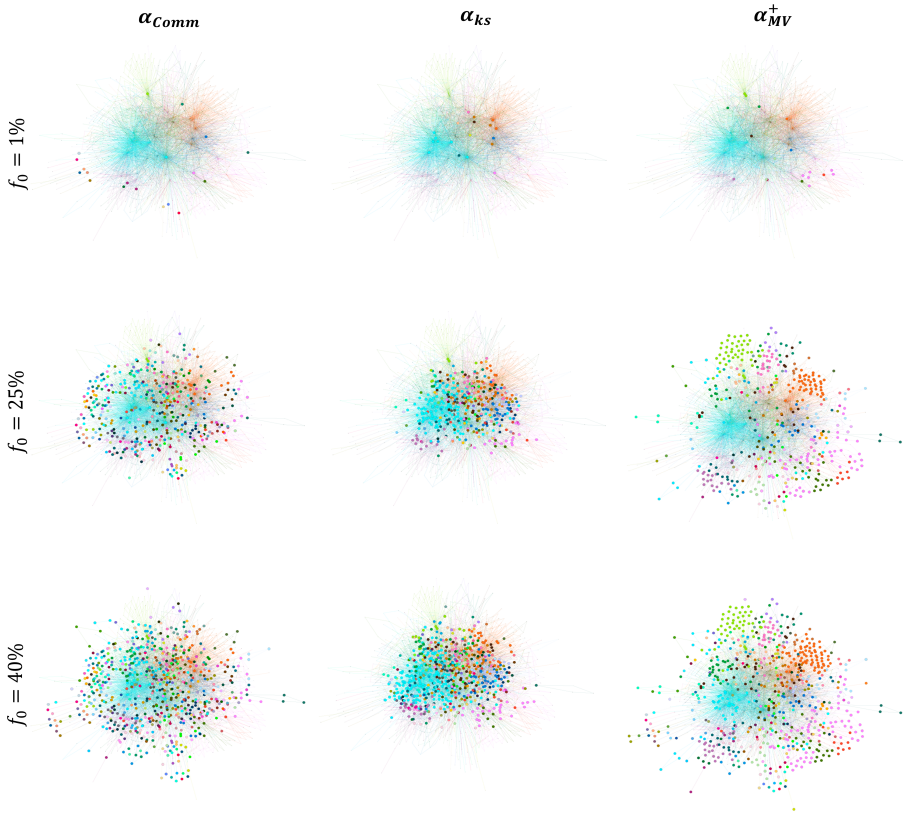


Figure 3.5: **Comparing the position of the top nodes in the Kegg Metabolic network (μ = 0.466).** The top nodes are chosen at a low budget availability (f_o = 1%), medium budget availability (f_o = 25%), and high budget availability (f_o = 40%). The bigger nodes in the left, middle, and right figures are the top nodes ranked by Comm Centrality (α_{Comm}), K-shell with Community (α_{ks}), and Modularity Vitality targeting hubs (α^{+}_{MV}), respectively.

In contrast, with K-shell with Community (α_{ks}), a measure that generally underperforms in these models, the nodes chosen are close to each other and embedded in the core of the network. Thus, the spreading virus or piece of information will die out before it reaches the peripherical areas in the network. Now, why do hub-like nodes targeted using Modularity Vitality targeting hubs (α^{+}_{MV}) at high budget availability yield the highest outbreak in the SI, SIR, and IC models? Referring back to Fig 3.5, when f_o = 40%, as we can see, the nodes are distributed even more than Comm Centrality (α_{Comm}) across all the regions in the network. Thus, it is normal to have a higher outbreak, as the virus/information would reach all the peripherical areas of the network and its core.

(2) Why is it more beneficial to target bridge-like nodes, regardless of the budget availability, in the SI, SIR, and IC diffusion models when the network has a strong community structure strength? We have seen that bridge-like nodes always perform well when the network has a strong community structure strength. The distinction lies in which community-aware centrality measure with a preference to bridge-like nodes yields the highest outbreak. The results show that Comm Centrality (α_{Comm}) generally performs best when the budget is limited. Conversely, when the budget availability is high, Modularity Vitality targeting bridges (α^{-}_{MV}) overcomes all the measures (see networks Ego Facebook and Facebook Politician Pages in Fig 3.4). To investigate these results, we visualize in Fig 3.6 the Facebook Politician Pages network, which has a strong community structure strength (μ = 0.111).

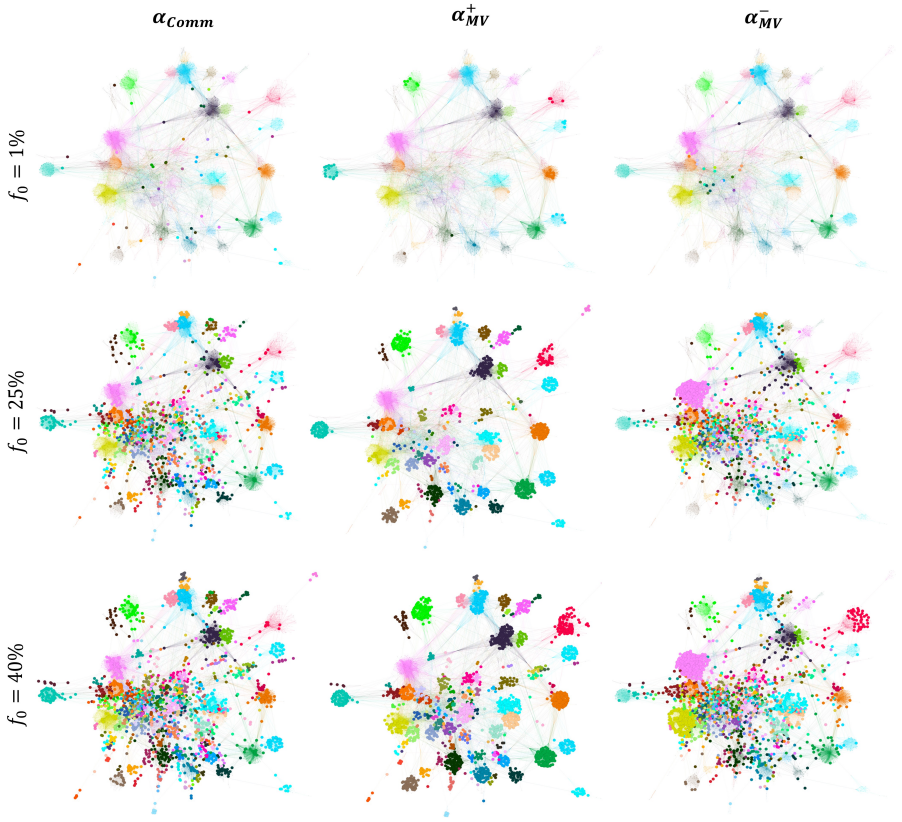


Figure 3.6: **Comparing the position of the top nodes in the Facebook Politician Pages network (μ = 0.111).** The top nodes are chosen at a low budget availability (f_o = 1%), medium budget availability (f_o = 25%), and high budget availability (f_o = 40%). The bigger nodes in the left, middle, and right figures are the top nodes ranked by Comm Centrality (α_{Comm}), Modularity Vitality targeting hubs (α^{+}_{MV}), and Modularity Vitality targeting bridges (α^{-}_{MV}), respectively.

For comparison purposes, the top nodes visualized are based on Comm Centrality (α_{Comm}), Modularity Vitality targeting hubs (α^{+}_{MV}), and Modularity Vitality targeting bridges (α^{-}_{MV}). As we can see, in this network with a strong community structure, when the budget is low (i.e., f_o = 1%) and medium (i.e., f_o = 25%), the top nodes ranked by Comm Centrality (α_{Comm}) are widespread between and across many communities. This spread indicates that the virus/information has many venues to further expand into, permitting a higher outbreak. As the budget increases to 40%, Modularity Vitality targeting bridges (α^{-}_{MV}) takes over. The nodes selected by it also spread across many regions in the networks, however, not to the extent of Comm Centrality (α_{Comm}), which reaches even the peripherical communities in the network. Indeed, Modularity Vitality targeting bridges (α^{-}_{MV}) focuses on bridges between communities and has a preference to target more nodes inside big communities rather than the peripherical areas. Thus, targeting bridges connecting communities and simultaneously focusing on big communities for higher outbreaks is more effective since small and peripherical communities cannot be leveraged as much as big communities if the budget is high and the network has a strong community structure. Thus, under the SI, SIR, and IC dynamics, choosing nodes inside and between the big communities diffuses the information more widely internally and externally. In contrast, small peripherical communities are isolated and do not have many pathways for the

virus/information to spread.

We also shed light on how Modularity Vitality targeting hubs (α^{+}_{MV}) behaves. At f_o = 0.40, Modularity Vitality targeting hubs (α^{+}_{MV}) does not target the community colored in fuchsia, the biggest community in the network. It does not since a node removed from a big and well-connected community will not change the network's modularity significantly. In contrast, a hub removed from a smaller community may shatter that community. Consequently, when ranked according to Modularity Vitality targeting hubs, these nodes would receive a higher score (α^{+}_{MV}). Thus, having big communities not targeted in a network with a strong community structure yields lower reachability of the virus/information. The behavior is the opposite in a network with a weak community structure. Since all the small communities surrounding a big community will be activated/infected, the infection/information has a higher probability of entering the big community as there are many pathways to enter it, causing an internal avalanche of infections/activations (see Kegg Metabolic in Fig 3.5 at f_o = 40%).

(3) Why is it more beneficial to target hub-like nodes at low budget availability in the LT model? Results reveal that the dynamics on the LT model contrast with that of the SI, SIR, and IC models. Indeed, bridge-like nodes are always preferred in the latter set of models when the budget is limited. However, with the LT model, under limited budget availability, hub-like nodes targeted by the Map Equation Centrality (α_{MapEq}) diffuse better the virus/information across the network. To understand why this phenomenon occurs, we visualize two structurally different networks, namely the Hamsterster and Facebook Politician Pages networks in Fig 3.7. In these two networks, the Map Equation Centrality (α_{MapEq}) shows good performances up to a specific budget. Suppose that a piece of information is circulating around a given community. If this community is well-connected (i.e., if it has a high internal density compared to its external connections), the piece of information will never enter it [99, 100]. This trend is even more pronounced when the nodes have a high threshold, even if the network has a weaker community structure.

The Map Equation Centrality (α_{MapEq}) overcomes this obstacle in the LT model by targeting nodes inside all the network communities and not around them. Because the Map Equation Centrality (α_{MapEq}) correlates with the node's intra-community links, the random walker has a higher chance of staying in nodes with a high internal degree. Thus, these nodes tend to be hub-like rather than bridge-like nodes, as seen in the two networks in Fig 3.7. For demonstration purposes, the nodes targeted by Community Hub-Bridge (α_{CHB}) and Comm Centrality (α_{Comm}) are also shown. These two measures perform poorly on the LT model when the budget is limited.

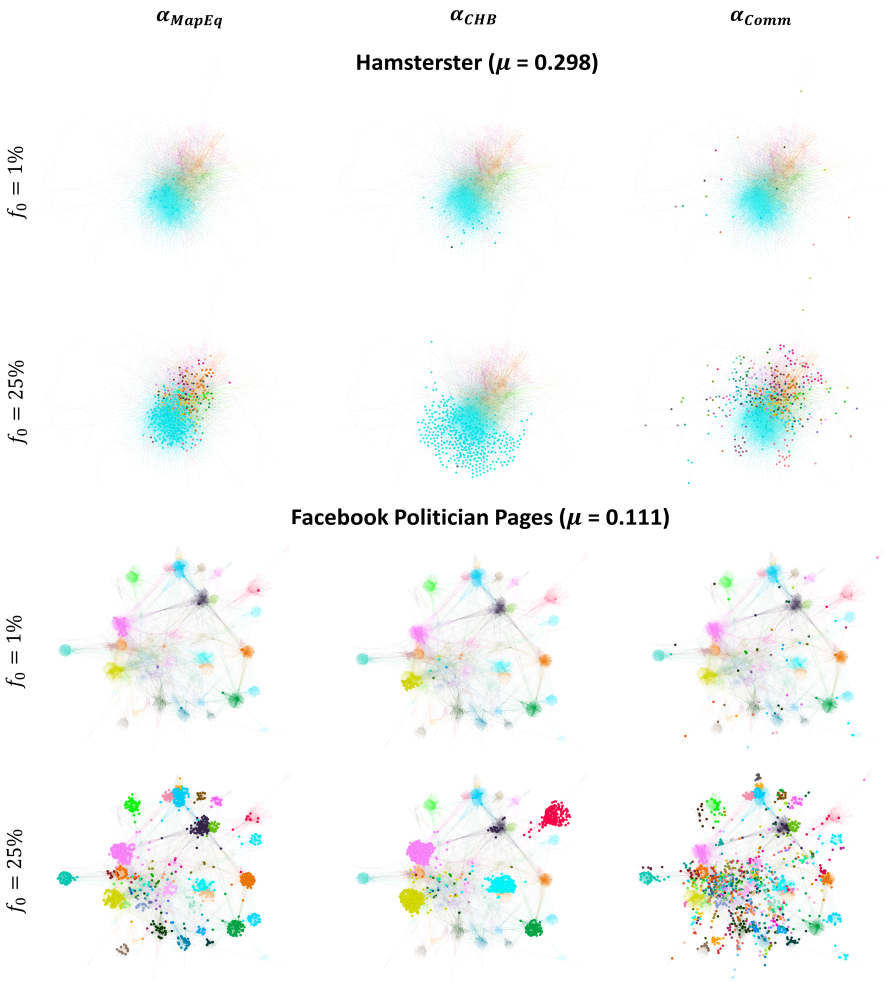


Figure 3.7: **Comparing the position of the top nodes in the Hamsterster and Facebook Politician networks.** The top nodes are chosen at a low budget availability (f_o = 1%) and medium budget availability (f_o = 25%). The bigger nodes in the left, middle, and right figures are the top nodes ranked by Map Equation Centrality (α_{MapEq}), Community Hub-Bridge (α_{CHB}), and Comm Centrality (α_{Comm}), respectively.

As we can see, Community Hub-Bridge (α_{CHB}) picks many nodes inside a few communities, missing many regions in the network. Concerning Comm Centrality (α_{Comm}), since the top nodes picked are more frequent between the communities rather than the inside as it has a preference for bridge-like nodes, this will not be enough at low budget availability for the piece of information to enter the tightly-knit communities [99, 100]. Therefore, given a low budget availability, ensuring a piece of information starts by nodes embedded in their communities such that these communities spread across all the network regions is the best approach for effective diffusion. If bridge-like nodes are targeted at low budget availability, the information will not be capable of entering high-density communities. Note that this behavior contrasts with the behaviors seen with the SI, SIR, and IC models. Indeed, in the latter set of models and at low budget availability, bridge-like nodes play the most influential role in diffusion since, in these models, bridge-like nodes have a higher chance to enter many communities and cause an avalanche of activations/infections.

(4) Why is it more beneficial to target both hub-like and bridge-like nodes simulta-

neously or bridge-like nodes only at high budget availability in the LT model?

At high budget availability, results show that either hub-like and bridge-like nodes targeted simultaneously using Modularity Vitality targeting hubs and bridges ($|\alpha_{MV}|$) or bridge-like nodes only using Comm Centrality (α_{Comm}) yield the highest outbreak. These trends also contrast with the ones found in the SI, SIR, and IC models. At high budget availability, the networks with the latter set of models showed good performance with either hub-like nodes targeted with Modularity Vitality targeting hubs (α^{+}_{MV}) or hub-like nodes targeted with Modularity Vitality targeting bridges (α^{-}_{MV}), depending on the community structure strength. This leads us to investigate why the LT dynamics also diverge when the budget availability is high. We visualize Ego Facebook and Facebook Politician Pages to depict the two trends regarding the outperformance of Modularity Vitality targeting hubs and bridges ($|\alpha_{MV}|$) in the former network and bridge-like nodes only using Comm Centrality (α_{Comm}) in the latter network in Fig 3.8. For comparison purposes, we also choose Community Hub-Bridge (α_{CHB}) to be represented.

Discussing first the Facebook Politician Pages network, we can see that both Modularity Vitality targeting hubs and bridges ($|\alpha_{MV}|$) and Comm Centrality (α_{Comm}) target nodes that all well-distributed, internally and externally, across all the communities in the network, unlike Community Hub-Bridge (α_{CHB}) which targets a limited number of communities. Since the communities in Facebook Politician Pages are not of equivalent sizes, Comm Centrality (α_{Comm}) yields a higher activation rate as the difference between Modularity Vitality targeting hubs and bridges ($|\alpha_{MV}|$) and Comm Centrality (α_{Comm}) is that the latter targets more nodes on the peripherical communities. In contrast, in the Ego Facebook network, since there is a smaller variance in the community size distribution, targeting hub-like and bridge-like nodes simultaneously using Modularity Vitality targeting hubs and bridges ($|\alpha_{MV}|$) assures that the diffusion will spread across the communities as small peripherical communities do not exist.

The question is, why do we observe such behavior in the LT model rather than the behavior seen with Modularity Vitality targeting hubs (α^{+}_{MV}) and Modularity Vitality targeting bridges (α^{-}_{MV}) in the SI, SIR, and IC dynamics. Targeting nodes inside the communities satisfies the dynamical conditions of the LT model for a higher activation rate [99, 100]. However, with a higher budget availability, bridge-like nodes also play a role since many are targeted. Subsequently, at high budget availability in the LT dynamics, enough hub-like and bridge-like will make the diffusion spread, rather than just targeting hub-like nodes or bridge-like nodes independently. Moreover, as we discussed previously, a major drawback for Modularity Vitality targeting hubs (α^{+}_{MV}) is that it misses hub-like nodes in big communities since they are easily replaced by others and subsequently they do not receive a high centrality score. Hence, it falls back in the LT dynamics as all communities should be targeted internally before externally for higher activation rates. We visualize the Ego Facebook network in Fig 3.9 with the top 40% nodes ranked by all the Modularity Vitality variants (i.e., α^{+}_{MV}, α^{-}_{MV}, and $|\alpha_{MV}|$) to show how Modularity Vitality targeting both hubs and bridges ($|\alpha_{MV}|$) is well suited for the LT dynamics as it assures internal diffusion and external diffusion by effectively utilizing the high budget availability. The red dashed lines highlight that two large communities in the network are not targeted by Modularity Vitality targeting hubs (α^{+}_{MV}) despite having a budget of f = 40%. In contrast, Modularity Vitality targeting both hubs and bridges ($|\alpha_{MV}|$) targets hub-like nodes inside all the communities and a set of bridges between them.

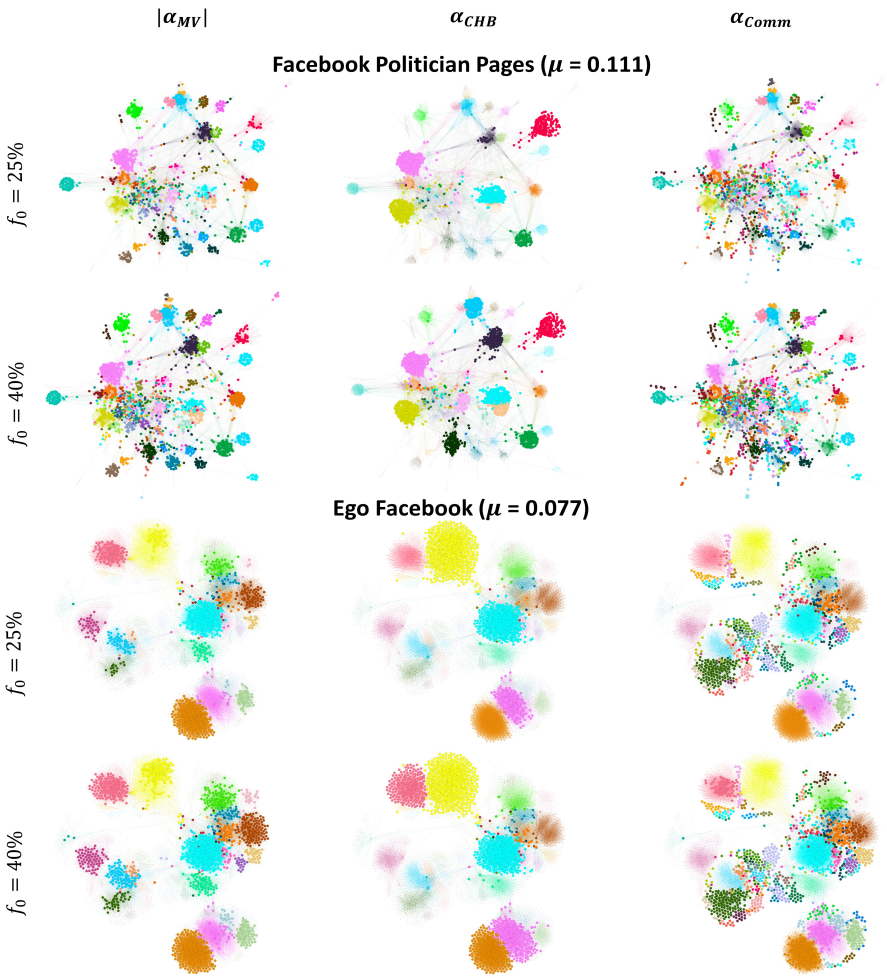


Figure 3.8: **Comparing the position of the top nodes in the Facebook Politician Pages and Ego Facebook networks.** The top nodes are chosen at a medium budget availability (f_o = 25%) and high budget availability (f_o = 40%). The bigger nodes in the left, middle, and right figures are the top nodes ranked by Modularity Vitality targeting hubs and bridges ($|\alpha_{MV}|$), Community Hub-Bridge (α_{CHB}), and Comm Centrality (α_{Comm}), respectively.

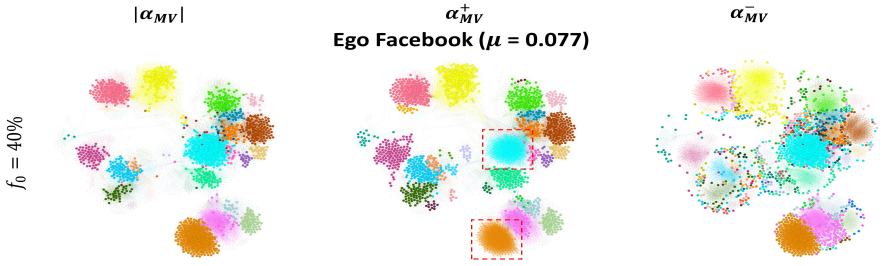


Figure 3.9: **Comparing the position of the top nodes in the Ego Facebook networks** The top nodes are chosen at a high budget availability (f_o = 40%). The bigger nodes in the left, middle, and right figures are the top nodes ranked by Modularity Vitality targeting hubs and bridges ($|\alpha_{MV}|$), hubs only (α^{+}_{MV}), and bridges only (α^{-}_{MV}), respectively.

(5) Comparing the dynamics of communities identified by Infomap and Louvain, why do the dynamics of the community-aware centrality measures the in the SI, SIR, and IC diverge when the budget is low to medium?

Results show that in the SI, SIR, and IC models, with a budget availability varying from low to medium, the nodes targeted by the community-aware centrality measures induce different dynamics when Infomap identifies the communities compared to the communities identified by Louvain. The measures having preferences for bridge-like nodes, namely Comm Centrality (α_{Comm}) and Community Hub-Bridge (α_{CHB}), underperform with Louvain as shown in the first three rows of Fig 3.10.

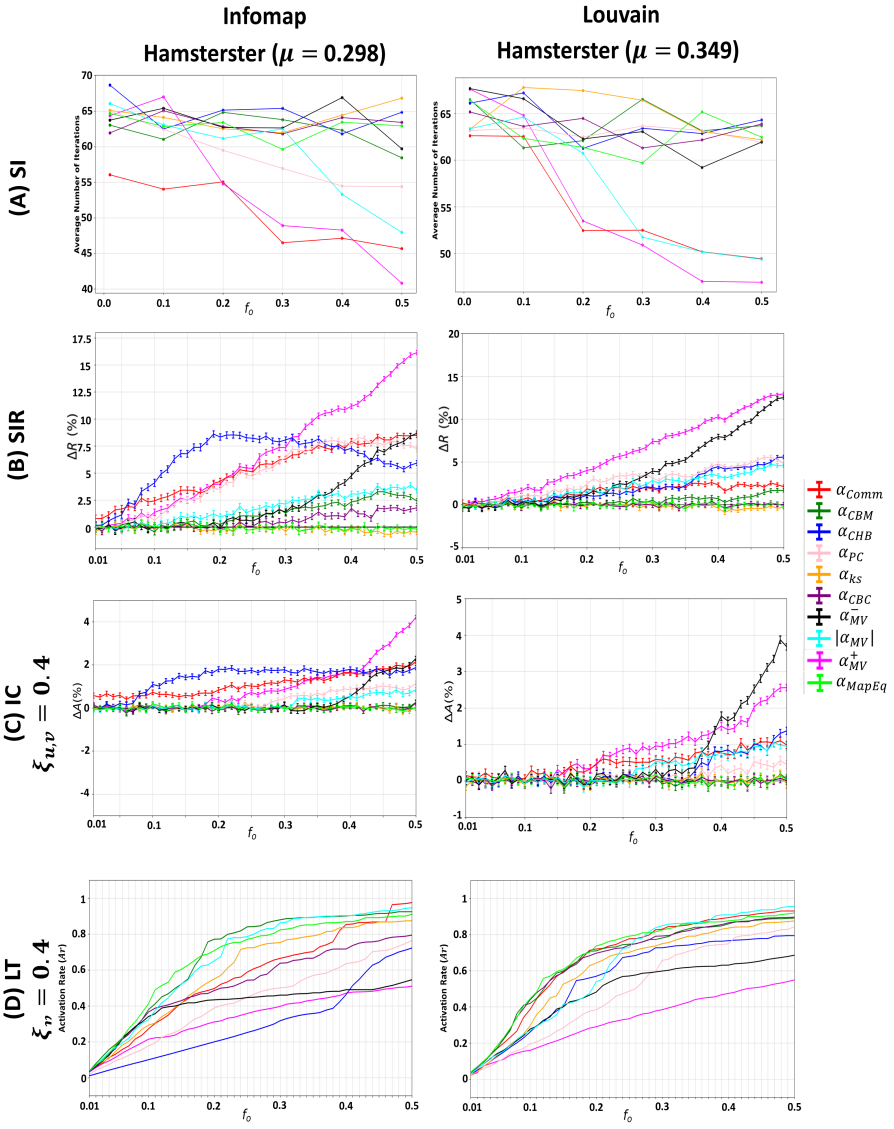


Figure 3.10: **Comparing the trends of the various dynamic models in Hamsterster with its communities identified by Infomap and Louvain.** The first, second, third, and fourth rows indicate the results of the (A) SI model, (B) SIR model, (C) IC model, (D) LT model.

We visualize the Hamsterster network with its communities identified by Infomap and

Louvain to clarify why this occurrence happens in Fig 3.11. We also show the histogram of the community size distribution in Fig 3.12 of the Hamsterster network using both Infomap and Louvain.

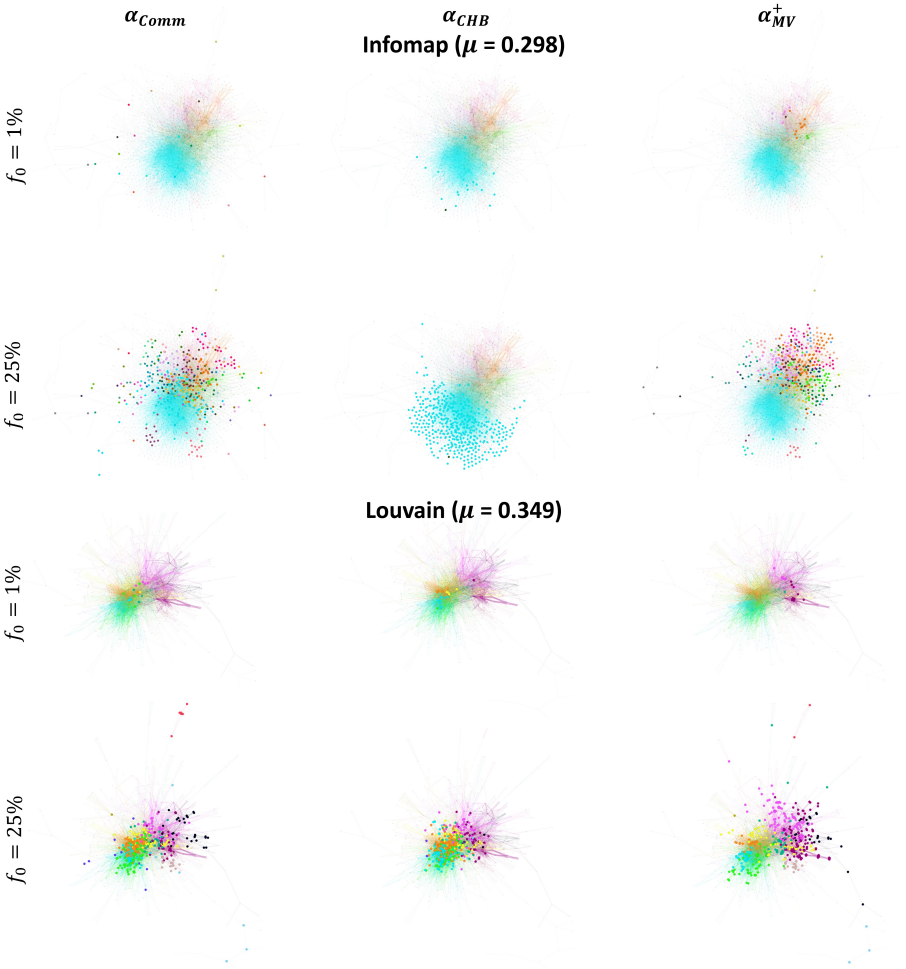


Figure 3.11: **Comparing the position of the top nodes in the Hamsterster network having its communities identified by Infomap and Louvain.** The top nodes are chosen at a low budget availability (f_o = 1%) and medium budget availability (f_o = 25%). The bigger nodes in the left, middle, and right figures are the top nodes ranked by Comm Centrality (α_{Comm}), Community Hub-Bridge (α_{CHB}), and Modularity Vitality targeting hubs (α^{+}_{MV}), respectively.

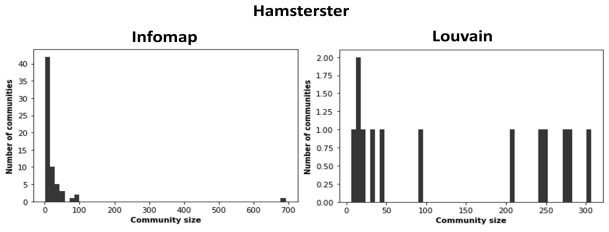


Figure 3.12: **Histograms of the community size distribution of the Hamsterster network.** Communities are identified by Infomap and Louvain.

Generally, Infomap yields high variance in the sizes of the communities with a power-law distribution. Louvain uncovers fewer communities with more uniform sizes having a lower variance. For instance, in the Hamsterster network, Infomap identifies 64 communities with a maximum size of 692 and a minimum size of 2. Conversely, Louvain uncovers 13 communities with a maximum size of 307 and a minimum size of 6. As Comm Centrality (α_{Comm}) exploits bridge-like nodes in all of the communities of the network, either small or large, having a more uniform size distribution with less variance diminishes Comm Centrality's return targeting. Indeed, the bridge-like nodes' frequency undoubtedly decreases with Louvain. As we can see in Fig 3.11, when the budget is f_o = 1%, and the communities are identified by Infomap, Comm Centrality's top nodes are well-distributed across the network in opposition to its top nodes selected when communities are determined by Louvain where they are situated in the core of the network. Similarly is the case at f_o = 25%.

(6) Comparing the dynamics on communities identified by Infomap and Louvain, why do the dynamics of the community-aware centrality measures in the LT diverge when the budget availability is high?

Results show that the dynamics of the LT model differ from that of the SI, SIR, and IC even when uncovering the communities with the Louvain algorithm. As the findings generally show, the differences in the outbreak can be seen when the budget is limited with the stated set of models. However, with the LT model, the differences are featured at high budget availability (see the last row of Fig 3.10). These differences accentuate how the LT dynamics differ from the remaining models. Yet, results show that bridge-like nodes also play a lesser role when Louvain identifies the communities. This is clear when f_o exceeds 0.47 in the LT dynamics of Hamsterster in the last row of Fig 3.10. At this given budget range, Comm Centrality (α_{Comm}) outperformed the remaining measures with Infomap. However, Modularity Vitality targeting hubs and bridges ($|\alpha_{MV}|$) outperformed with Louvain. In addition, we note that Map Equation Centrality (α_{MapEq}) and Community-based Centrality (α_{CBC}) show superior performances with Louvain as these measures prioritize hub-like nodes, which are pervasive with Louvain as compared to Infomap.

3.6/ CONCLUSION

Modeling complex network dynamics is a major breakthrough in describing and understanding the real world. Researchers from various disciplines, such as sociology, epidemiology, and physics, have developed diffusion models deemed to be interdisciplinary in nature. These diffusion models differ mainly in their underlying conditions and states as the dynamic process begins in a given network. In the vast data era we live in, a myriad of unique topological characteristics characterizes networks. One of the prominent characteristics is the network's community structure. Indeed, the community structure affects any diffusive phenomena on the network.

That being said, finding the most important nodes that play a role in accelerating or inhibiting a diffusion phenomenon within and across these communities is of utmost importance. Community-aware centrality measures acknowledge the network's community structure and aim to identify key nodes accordingly. Some measures prioritize hub-like nodes, while others prioritize bridge-like nodes. Still, the aim at the end is to maximize the diffusion (or inhibit it) under any dynamic model in a network.

Numerous community-aware centrality measures and diffusion models have been proposed in the literature. This chapter investigates the interplay between the diffusion dynamics, the community-aware centrality measures, and the network's topological characteristics. More specifically, we analyze how the diffusive power of nodes selected based on various community-aware centrality measures changes with respect to the model and the network at hand. Four diffusion models have been simulated, starting with a set of initial nodes based on the community-aware measures under study on synthetic and real-world networks. The diffusion models studied are the Susceptible-Infected (SI), Susceptible-Infected-Recovered (SIR), Independent Cascade (IC), and Linear Threshold (LT) models.

Results show that the strength of the community structure and budget availability significantly impact how diffusion spreads. In addition, the SI, SIR, and IC dynamics show a convergent behavior, while the LT dynamics diverge within a given community structure strength and budget availability. By controlling the community structure strength in synthetic networks, we observed that the community-aware centrality measures are more profitable in networks with a strong community structure strength. With real-world networks with a strong community structure under the SI, SIR, and IC dynamics, bridges are always preferred regardless of the budget. With the LT dynamics, hub-like nodes are preferred when the budget is limited or high. However, when the budget increases, hub-like and bridge-like nodes are preferred. In networks with a weak community structure, with the SI, SIR, and IC dynamics, bridge-like nodes are preferred, then distant hub-like nodes take over at high budget availability. However, with the LT dynamics, hub-like nodes are preferred at a low budget, while more interlinked nodes with hub-like nodes are preferred from medium to high budget availability. We also analyzed the impact of the community detection algorithm, and results showed that in the SI, SIR, and IC dynamics, the performance of the measures changes when the budget is limited. In contrast, with the LT dynamics, differences are seen when the budget availability is high. The differences between the diffusion models, mainly seen at a limited budget availability, is credited to the fact that the conditions in the SI, SIR, and IC models are well suited to select bridge-like nodes as it is easier for the virus/piece of information to circulate from one community to another compared to the LT model. Indeed, if the virus/piece of information is initiated in the well-connected communities under the LT dynamics, the community will never be part of the occurring diffusive phenomenon.

The extensive experiments shed light on how the diffusion dynamics, the position of the nodes initially activated, the network's community structure strength, and the budget availability are interconnected. Given the knowledge of one or the other, one can choose the suitable measure for running the most effective diffusion on the network.

4

A Community-Aware Ranking Scheme

4.1/ Introduction

Many real-world systems, including transportation, social, technological, infrastructural, information, and biological systems, are complex. Influential nodes in these systems play a critical role in the structure and dynamics of the network [107]. Identifying the most influential nodes in these networks is a major issue. Indeed, it allows for conducting specific optimization tasks, such as controlling, minimizing, or maximizing a diffusion process. This issue is mainly related to centrality measures [7]. These measures extract diverse information from the network to quantify its importance. For instance, betweenness centrality allows identifying genes related to heart attacks [108]. Other applications include hindering epidemic outbreaks [109], augmenting the effectiveness of marketing campaigns on social media [110], enhancing the resiliency of infrastructural networks [111], and many other [112–114].

Centrality measures can be signed (positive or negative) or unsigned (positive and negative). In the first case, one ranks nodes in descending order, and a fraction of the top nodes are selected to conduct a specific optimization task. In the second case, one can have a multitude of ranking schemes. One can combine positive and negative ranks, take a fraction of both positions, or convert negative values to positive values and take the aggregate ranks. In either case, one can have two general ranking schemes, strong and weak. In the former, one selects the most critical nodes first. In the latter, one chooses the less important nodes [115]. Although centrality measures provide an effective way to rank nodes, several challenges exist. The first challenge is that several centrality measures may underestimate the influence of specific nodes depending on the network's structure [116]. The second challenge is that many nodes with high centrality may be neighbors [117]. Thus, targeting these nodes for diffusion or immunization is inefficient because one uses the resources locally, ignoring large parts of the network. The third challenge is which ranking schemes and/or combination criteria are ideal for a network with specific topological features.

To address these challenges, we propose a ranking method that considers the community structure. Communities are widely present in real-world networks, consisting of tightly connected and coherent sets of nodes with only a few links with nodes outside their group [1]. The community structure of a network is known to impact its underlying dynamics [118, 119], and recent studies highlight the advantages of using this structure

to identify influential nodes [20–27]. The proposed ranking strategy exploits this precious information. It is simple yet effective and applicable to all centrality measures. Given a centrality measure, the most straightforward ranking strategy targets the top nodes independently of the community structure, if any. Instead, we propose to rank the nodes based on their importance in their communities. First, we select the most central nodes in each community. We order these nodes in decreasing order of their community size. Then we move to each community's next most central node and adopt the same ordering strategy. We iterate this process until we reach the given budget of nodes to rank. This approach naturally selects distant nodes in each community.

To evaluate the proposed ranking strategy, we report a series of experiments on synthetic and real-world networks using a set of six classical centrality measures using the SIR epidemic model [120]. We categorize these centrality measures into three groups, namely neighborhood-based (Degree and Maximum Neighborhood Component), path-based (Betweenness and Closeness), and iterative refinement-based (Katz and PageRank). Experiments on synthetic networks investigate the impact of various network parameters on the proposed ranking strategy. Indeed, one can control the community structure strength, the community size distribution, and the degree distribution. Real-world networks include infrastructural, social, biological, citation, word, and collaboration networks with unknown community structures. Therefore, we uncover the communities using two community detection algorithms to assess the proposed strategy's consistency linked to community structure variations. Results show that the community-aware ranking strategy is more effective than the classical ranking by descending order of the centrality measure. The main advantages of the proposed method are threefold:

1. It applies to all types of centrality measures in all networks (undirected/directed and unweighted/weighted).

2. It naturally selects distant nodes to expand any diffusion phenomena based on any given budget.

3. Its complexity depends on the centrality measure computed.

4.2/ PROPOSED RANKING STRATEGY

Centrality measures attempt to measure the level of influence nodes have in a network. Nodes are typically ranked in decreasing order of influence, with those having the highest centrality considered to be strategically positioned within the network. As a result, targeting these nodes for optimization tasks will likely result in favorable outcomes. However, in real-world networks, these nodes may be close, which can harm the effectiveness of dissemination strategies. Consider an immunization scenario in an epidemic process. Prioritizing the immunization of neighbors, even if influential, may prevent protecting vast areas of the network. To address this problem, we suggest a community-aware ranking scheme that considers the network's community structure to select distant influential nodes.

ALGORITHM

The proposed ranking strategy targets influential nodes spreading across communities in the network. It applies to any centrality measure. First, one computes the centrality of the nodes. Second, one targets top nodes community by community. Such a strategy prevents the concentration of influential nodes in the same network area. The targeted nodes are naturally more dispersed.

Algorithm 1 Community-aware ranking scheme

Input: Graph $G(V, E)$, Centrality measure β, Sorted community set C, Budget B
Output: List of ranked nodes L

$D \leftarrow \varnothing$ ▷ Compute the centrality of each node
for each $i \in V$ **do**
    $D[i] \leftarrow \beta(i)$
for each $c_{l,l\in\{1,2,\ldots,|C|\}} \in C$ **do** ▷ Sorting the nodes inside their communities
    for each $i \in c_l$ **do**
        $D_{c_l} \leftarrow D[i]$
    $D_{c_l} \leftarrow Sort(D_{c_l})$
while $B \neq 0$ **do** ▷ Extract sorted list of nodes till budget is reached
    for each D_{c_l} and $i \in D_{c_l}$ **do**
        if $D_{c_l} \neq \varnothing$ **then**
            $v \leftarrow D_{c_l}.pop(i)$
            $L.append(v)$
            $B \leftarrow B - 1$

Doing so makes it more likely that any diffusion process spreads more uniformly in the network than in the case where targeted nodes by a centrality measure in a descending order ranking scheme are close to each other. Note that we assume the communities are sorted from the biggest to the smallest, with ties decided at random. Also, note that the maximum budget is the size of the network. The pseudocode is provided in Algorithm 1.

TOY EXAMPLE

Figure 4.1 illustrates the proposed ranking method on a toy example. The network contains 22 nodes and three communities in this example. Suppose the maximum budget is three nodes out of the whole network. We consider Degree and Betweenness centrality as measures of influence. Tables 3 and 4 in Appendix I report the centrality values and the corresponding ranks using the descending order and the proposed approach. Based on the descending order ranking scheme of the Degree centrality, presented on top of Fig 4.1A, we can see that the highest degree nodes (nodes 1, 4, and 5) belong to the same community *C1*. Similarly, the nodes with the highest Betweenness centrality (nodes 13, 14, and 15) are all located in the same community *C2*, presented at the bottom of Fig 4.1A.

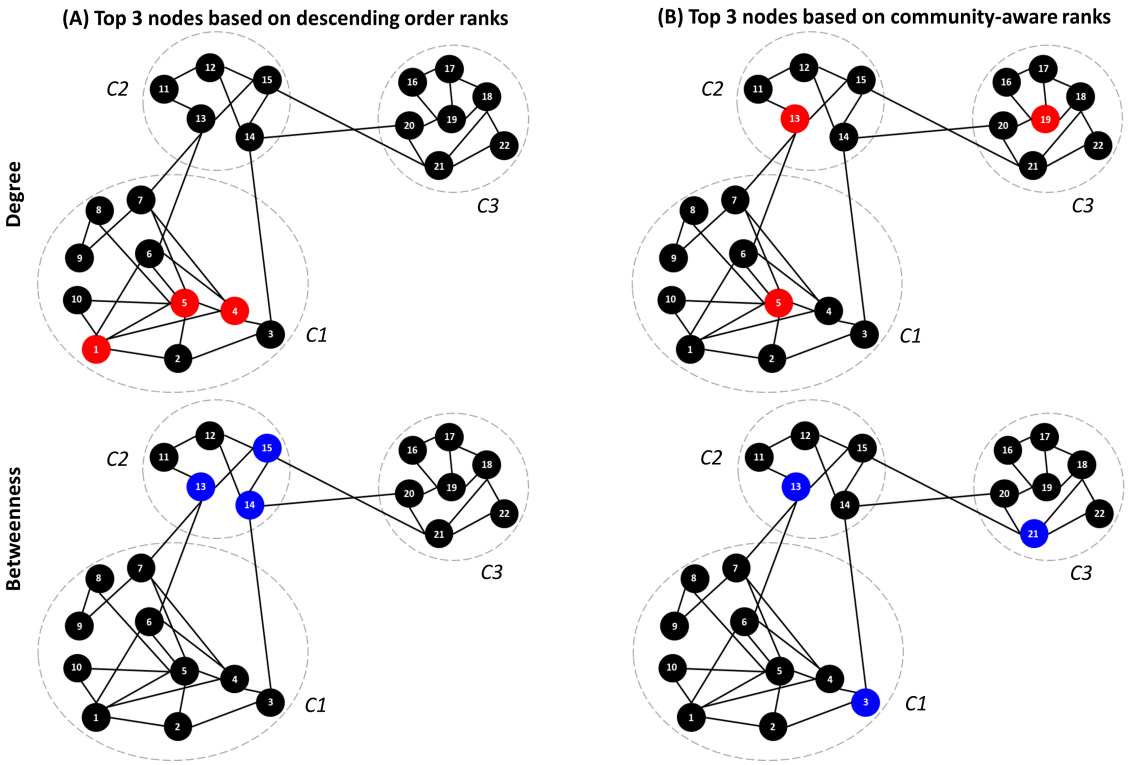


Figure 4.1: **Illustrating the behavior of the descending order ranking scheme and the community-aware ranking scheme.** The nodes chosen are the top 3 nodes based on the Degree centrality (colored in red) and the Betweenness centrality (colored in blue).

In contrast, the proposed community-aware ranking scheme selects the highest degree node in each community, presented on top of Fig 4.1B. Indeed, node 5 is picked from community *C1*, node 13 is picked from community *C2*, and node 19 is picked from community *C3*. Results with the Betweenness centrality are similar, as presented at the bottom of Fig 4.1B. Instead of targeting the top three nodes in the same community *C2*, the proposed ranking approach selects nodes with the highest Betweenness centrality in each community. More precisely node 3 in *C1*, node 13 in *C2*, and node 21 in *C3*. Note that ranks of nodes with the same centrality value in a community are chosen randomly.

One of the main drawbacks of the classical descending order ranking scheme is ignoring the network's community structure. From the diffusion perspective, if targeted nodes diffusing a piece of information or a virus are very close, the diffusion dies out before spreading across the other communities. On the contrary, the proposed ranking approach naturally selects the most influential nodes in their community. Indeed, the proposed ranking scheme favors nodes from all the dense parts of the network rather than specific communities.

4.3/ SYNTHETIC NETWORKS

We investigate synthetic networks using the LFR benchmark [97].

INFLUENCE OF THE COMMUNITY STRUCTURE STRENGTH

This experiment aims to investigate the influence of the community structure strength on the performance of the ranking strategies (the descending order ranking scheme and the proposed community-aware ranking scheme). The mixing parameter (μ) is tuned to cover a wide range of community structure strengths. It spans from very strong to very weak (μ = 0.05, 0.10, 0.20, 0.40, 0.70). Remember that a low value means few links between communities, indicating a strong community structure. In contrast, high value corresponds to networks with many links between communities, indicating a weak community structure. For the sake of brevity, we only show the outcomes of the most significant situations.

Fig 4.2 shows the relative difference in the outbreak size (ΔR) as a function of the fraction of initially infected nodes of the six investigated centrality measures (Degree, Maximum Neighborhood Component, Betweenness, Closeness, Katz, PageRank) with a strong (μ = 0.05), medium (μ = 0.40), and weak (μ = 0.70) community structure strengths. The remaining parameters, including the community size (θ) and degree distribution (γ) exponents, are fixed at 2.7. The outbreak size (ΔR), represented by the red curve, is the difference between the number of nodes recovered after an initial set of nodes ranked based on the community-aware ranking scheme is infected and another initial set of nodes infected ranked based on the classical descending order ranking scheme. Thus, it represents a measure of performance of the community-aware ranking scheme. Positive values indicate that the proposed ranking scheme performs better (see Appendix I for details about calculating ΔR).

In networks with a strong community structure (μ = 0.05), the community-aware ranking scheme consistently outperforms the classical descending order ranking scheme for all the centrality measures under study. The gain reaches 24% for Katz centrality at a fraction of initially infected nodes (f_o) of 0.20, followed by 22% for Degree, MNC, and Closeness centrality. The performance of these measures is consistent from a fraction of initially infected nodes of 0.10 till 0.25, then they decline. Closeness centrality slightly declines, showing a ΔR of 14% at f_o = 0.50, followed by Katz centrality with 8%, then Degree and MNC obtaining a ΔR of 6%. Betweenness and PageRank are the less performing measures under the community-aware ranking scheme. The maximum gain for Betweenness is 12.5% at f_o = 0.12, and for PageRank is 16.5% at f_o = 0.9. After a peak, performance declines, reaching a gain of 2% at f_o = 0.50.

In networks with a medium community structure (μ = 0.40), the community-aware ranking scheme of all the centrality measures still performs better than the classical descending order ranking scheme. When the fraction of initially infected nodes f_o is low (i.e., between 0.01 and 0.05), the gain for all the centrality measures is low, reaching a maximum of 1%. As the fraction of initially infected nodes increases, the performance of the community-aware ranking scheme also increases until it reaches a plateau or barely changes. For example, the relative difference in the outbreak size of Degree centrality increases from f_o equaling 0.10 to 0.25, going till ΔR = 6.5%, and then it hardly changes. MNC, Betweenness, Closeness, Katz, and PageRank show similar behavior with ΔR reaching a maximum between 5% and 10%.

In networks with a weak community structure (μ = 0.70), when the fraction of initially infected nodes is between 0.01 and 0.10, the relative difference in the outbreak size (ΔR) alternates between -1% and +1%. After that, it increases to a maximum of ΔR = 3%

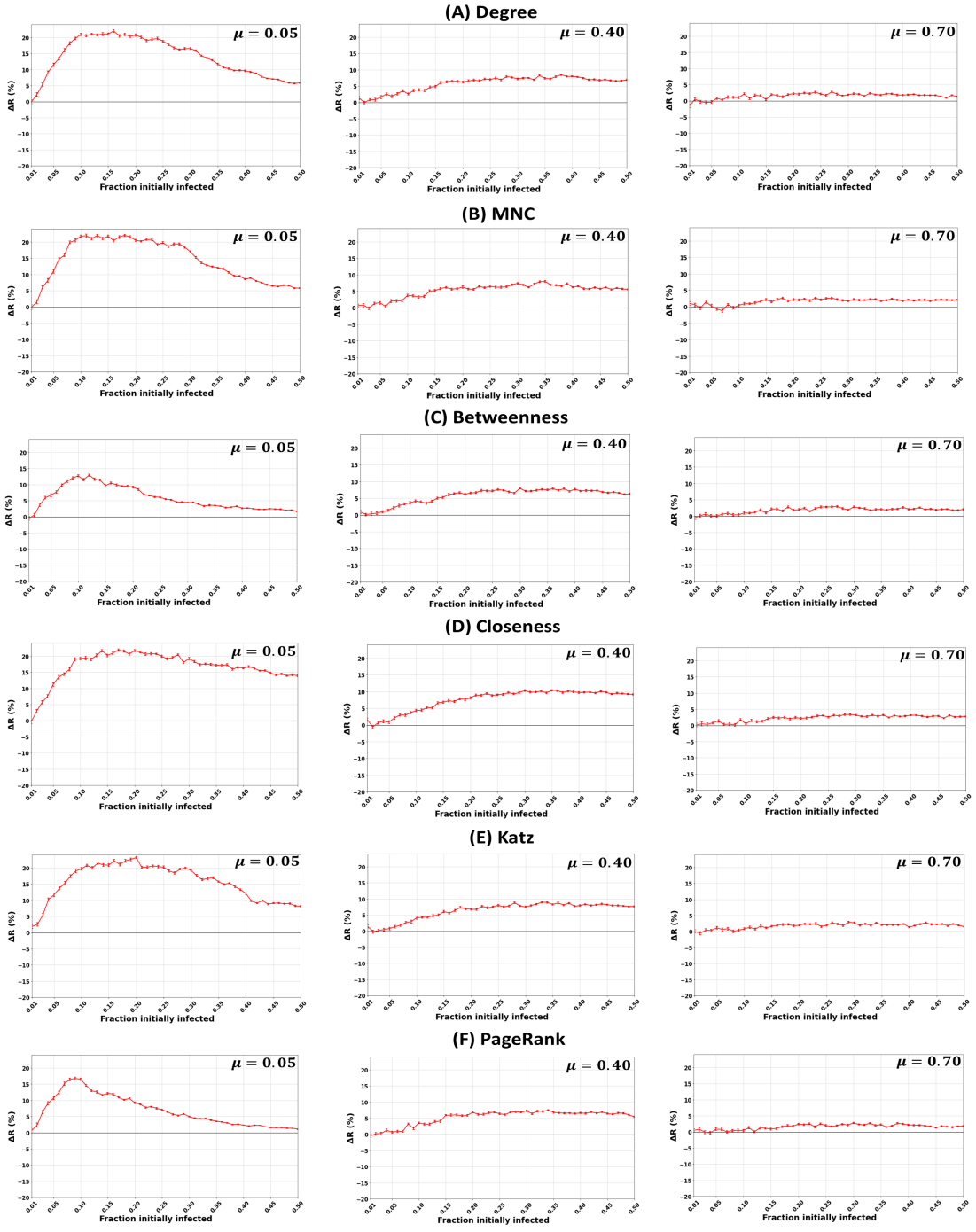


Figure 4.2: **Impact of the community structure strength (μ) in synthetic networks.** The figures represent the relative difference of the outbreak size (ΔR) as a function of the fraction of initially infected nodes. The red curve indicates the relative performance difference of the community-aware ranking strategy with the descending order ranking for the six centrality measures under test. The mixing parameter (μ) varies while the other parameters, including the community size distribution exponent ($\theta = 2.7$) and the degree distribution exponent ($\gamma = 2.7$), are fixed.

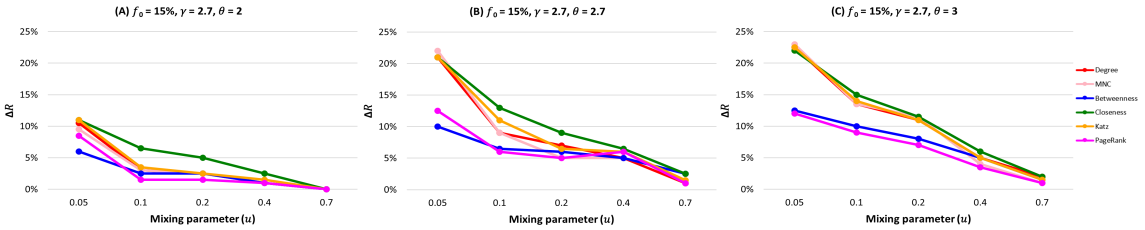


Figure 4.3: **The relative difference of the outbreak size (ΔR) as a function of the mixing parameter (μ) when fraction of initially infected nodes (f_o) equals 0.15.** The color of the curve represents the centrality measures under study. (A) Synthetic networks with degree distribution γ = 2.7 and community size distribution θ = 2. (B) Synthetic networks with degree distribution γ = 2.7 and community size distribution θ = 2.7. (C) Synthetic networks with degree distribution γ = 2.7 and community size distribution θ = 3.

and shows a plateau. One can expect these results. Indeed, the frontier between weak community structure and no community structure is thin.

We fix the fraction of initially infected nodes at 0.15 for all the centrality measures in Fig 4.2 and plot the relative difference in the outbreak size (ΔR) as a function of the mixing parameter (μ) as shown in Fig 4.3B. As the community structure gets weaker (i.e., from μ = 0.05 to μ = 0.7), the performance of the community-aware ranking scheme starts declining. Moreover, one can differentiate between the centrality measures' effectiveness. Closeness is the best-performing centrality measure, followed by Katz, Degree, and MNC. In contrast, Betweenness and PageRank perform poorly. However, all the measurements show a higher relative difference in the epidemic outbreak size than the classical descending order ranking scheme.

These results show that the community-aware ranking scheme is more effective in networks with a strong community structure. Indeed, in a strong community structure, communities are so well-separated that one can consider them independent subnetworks with their topological characteristics. In turn, targeting the most influential nodes in each community leads to a higher spreading, ensuring that the diffusion reaches all communities. As the community structure gets weaker, the performance of the community-aware ranking scheme decreases. Since the community structure is not well defined, the network is barely distinguishable from the one without a community structure. However, even in the worst-case scenario, the community-aware ranking scheme still is more effective than the classical descending order ranking scheme.

Influence of the Community Size Distribution

This investigation aims to analyze the impact of the community size distribution on the community-aware ranking scheme. One can tune the power-law community size distribution exponent (θ) in the networks generated by the LFR. In this study, we evaluate two values representing extreme cases. In the first case with θ = 2, few small communities coexist with large communities with a large variance in community sizes. In the second case, with θ = 3, more communities of equivalent sizes coexist, and the variance in the community sizes is minor. There are more communities in the second case than in the first case. Table 5 in Appendix I reports the number of communities of each generated network, along with the minimum and maximum size of the communities. Note that we

also perform tests with $\theta = 2.7$. However, no significant differences were compared to θ = 3.

Fig 4.4 shows the relative difference in the outbreak size (ΔR) as a function of the fraction of initially infected nodes for the Degree and Katz centrality. The community size distribution exponent θ equals 2 (panel A) and 3 (panel B). The other fixed parameters include the mixing parameter (μ = 0.05) and the degree distribution exponent (γ = 2.7).

When θ = 2 (Fig 4.4A), the networks contain a few small communities coexisting with much larger ones. The relative difference of the outbreak size (ΔR) increases as the fraction of initially infected nodes increases, reaching a maximum of 11% for Degree centrality and 12% for Katz centrality for the community-aware ranking scheme. Then, ΔR barely varies when the fraction of initially infected nodes f_o ranges from 0.10 and 0.25. After that, it gradually decreases, reaching 3.5% and 4% for both centralities, respectively, when f_o = 0.50.

When θ equals 3 (Fig 4.4B), many small communities of comparable sizes and a few large ones exist. The performance of the community-aware ranking scheme for Degree centrality increases, reaching a maximum of 24% gain in terms of ΔR. Then it gradually decreases until it reaches 5.1% gain when the fraction of initially infected nodes is 0.50. Katz centrality exhibits similar behavior. The relative outbreak size increases as the fraction of initially infected nodes increases, reaching a maximum gain of 24%. It decreases until it reaches a gain of 8% when the fraction of initially infected nodes equals 0.50.

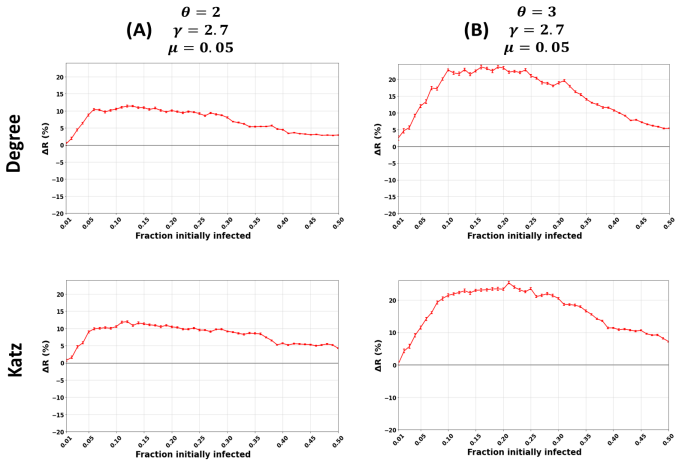


Figure 4.4: **Impact of the community size distribution exponent (θ) in synthetic networks.** The figures represent the relative difference of the outbreak size (ΔR) as a function of the fraction of initially infected nodes. The red curve indicates the relative performance difference of the community-aware ranks of the Degree and Katz centrality measures compared to the descending order ranks. The community size distribution exponent (θ) varies while the other parameters, including the mixing parameter (μ = 0.05) and the degree distribution exponent ($\gamma = 2.7$), are fixed.

Fig 4.3 reports the relative outbreak size (ΔR) for all the centrality measures in synthetic networks with a community size distribution exponent (θ) spanning from 2 (Panel A) to 3 (Panel C). The fraction of initially infected nodes (f_o) is fixed at 15%. When networks have a large difference in the sizes of the communities leading to fewer communities (θ = 2), the gain in ΔR of the community-aware ranking scheme ranges from 11% as a maximum at

μ = 0.05. It decreases, reaching 0% when μ = 0.70. On the contrary, when the networks have many small communities with fewer larger ones leading to many communities, ΔR for Degree, MNC, Closeness, and Katz reach a gain of 23% and a gain of 13% and 12.7% for Betweenness and PageRank, respectively. As the community structure gets weaker, ΔR decreases to a minimum of 1% for PageRank centrality and a maximum of 2.3% for Closeness centrality.

Results indicate that when the network contains a few large communities, the community-aware ranking scheme is not as effective as in networks with many communities of smaller sizes. It is reasonable since when huge communities coexist with a few small communities, the large communities will make up most of the network. When one picks the top influential nodes from each community in the first iteration, the nodes picked in the second iteration inside the large communities will likely be next to each other. Indeed, when there are substantial communities, there are few communities overall. Thus, infecting the most influential nodes in the same neighborhood is less effective than covering many communities of comparable sizes with the community-aware ranking scheme.

INFLUENCE OF THE DEGREE DISTRIBUTION

In this experiment, we investigate the effect of the degree distribution on the performance of the community-aware ranking scheme. The degree distribution exponent (γ) is tunable in the LFR model. Studies have shown that real-world networks are scale-free, with a degree distribution exponent in the range of 2 and 3 [104, 121]. Consequently, we test these two values. We also set $\gamma = 2.7$, but there are no significant differences compared to γ = 3. When γ equals 2, the network's structure resembles a hub-and-spoke network [105]. When γ equals 3, the network's structure is more similar to a random network where more nodes would have a comparable frequency of neighbors. Since the LFR model also generates networks with a community structure, the nodes inside the communities have equivalent sizes while ensuring the community structure is maintained [97].

Fig 4.5 shows the relative difference in the outbreak size (ΔR) as a function of the fraction of initially infected nodes for the Degree and Katz centrality. The degree distribution exponent γ equals 2 (panel A) and 3 (panel B). We fix all the other parameters, including the mixing parameter (μ = 0.05) and the community size distribution exponent (θ = 2.7).

When γ = 2 (Fig 4.5A), generating networks with a hub-and-spoke structure, the relative difference of the outbreak size (ΔR) of both Degree centrality and Katz centrality under the community-aware ranking scheme escalates quickly from a fraction of initially infected nodes (f_o) amounting to 0.01 till 0.10, reaching a maximum of 24%. ΔR stays in this range between 20% and 24% from f_o = 0.11 till f_o = 0.27 for Degree and f_o = 0.30 for Katz. After which ΔR starts to decrease, reaching 6.5% for Degree and 8.5% for Katz at f_o = 0.50.

When γ = 3, the degree distribution of the communities of the generated networks is more random than average. One can observe that both Degree centrality and Katz centrality perform similarly according to the relative difference of the outbreak size (ΔR). Compared to the descending order ranking scheme, the gain of the community-aware ranking scheme reaches 19% and barely changes till f_o = 0.25. Then it gradually decreases, reaching ΔR = 5% at f_o = 0.50.

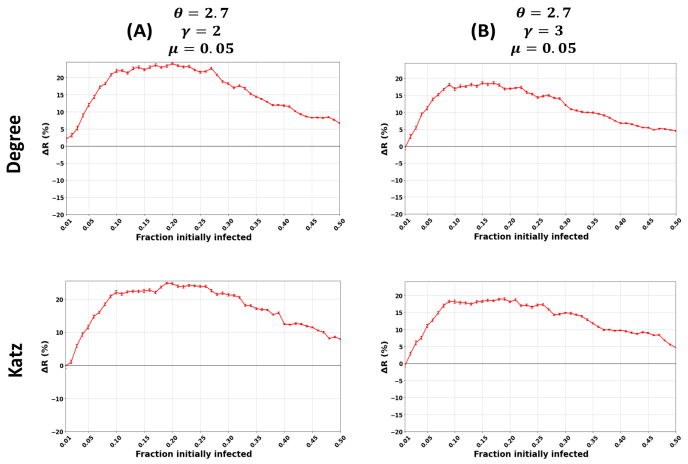


Figure 4.5: **Impact of the degree distribution exponent (γ) in synthetic networks.** The figures represent the relative difference of the outbreak size (ΔR) as a function of the fraction of initially infected nodes. The red curve indicates the relative performance difference of the community-aware ranks of the Degree and Katz centrality measures compared to the descending order ranks. The degree distribution exponent (γ) varies while the other parameters, including the mixing parameter (μ = 0.05) and the community size distribution exponent (θ = 2.7), are fixed.

Fig 4.6 reports the differences in the relative outbreak size (ΔR) for all the centrality measures in networks with a degree distribution exponent (γ) spanning from 2 (Panel A) to 3 (Panel C). The fraction of initially infected nodes (f_o) equals 15% When networks are similar to a hub-and-spoke structure (Fig 4.6A) with a strong community structure (μ = 0.05), centrality measures under the community-aware ranking scheme always show a higher relative outbreak size difference (ΔR).

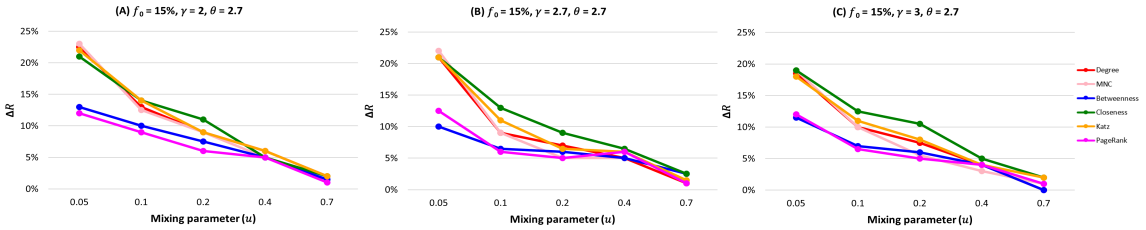


Figure 4.6: **The relative difference of the outbreak size (ΔR) as a function of the mixing parameter (μ) when fraction of initially infected nodes (f_o) equals 0.15.** The color of the curve represents the centrality measures under study. (A) Synthetic networks with degree distribution γ = 2 and community size distribution θ = 2.7. (B) Synthetic networks with degree distribution γ = 2.7 and community size distribution θ = 2.7. (C) Synthetic networks with degree distribution γ = 3 and community size distribution θ = 2.7.

However, one can consider two categories. The first, including Degree, MNC, Closeness, and Katz, exhibit gains ranging from 21% and 23.5%. The second involving Betweenness and PageRank, obtains a gain of around 13%. As we shift to a more random-like structure (γ = 3) in Fig 4.6C, categorizing the centrality measures observed at γ = 2 remains the same. However, for the first category, the gain decreases between 17% and 18%. The second category exhibits a gain of around 11%. As the community structure weakens, the difference in the outbreak size becomes less pronounced. Nevertheless,

the community-aware ranking scheme always performs better than the descending order ranking scheme.

Even though the differences are not as pronounced compared to the variation in the community size distribution, results show that when the communities of the generated networks are more random-like, the performance of the community-aware ranking scheme slightly decreases. Since more nodes have a comparable number of connections internally in a random-like structure, they may have similar centrality values. In turn, the community-aware ranking scheme may be prone to selecting nodes of the same influence inside their communities, saturating the diffusion spread. On the contrary, in a network with well-separated communities such as the hub-and-spoke structure, a community-aware ranking scheme can distinctively pick influential nodes in their communities that are naturally not close to each other due to the hub-and-spoke structure. It results in a higher diffusion to more isolated areas that the descending order ranking scheme cannot reach.

4.4/ REAL-WORLD NETWORKS

We also investigate the community-aware ranking scheme on 33 real-world networks covering various domains (i.e., infrastructural, biological, social, collaboration, and ecological). Since their community structure is unknown, we uncover it using two community detection algorithms: Infomap [106], and Louvain [87]. It allows us to check the impact of the community structure variations on the consistency of the community-aware ranking scheme.

SPREADING POWER OF THE PROPOSED METHOD

Since the community structure strength is a significant feature influencing the performance of the proposed community-aware ranking strategy, we classify the networks into three categories. The categories cover networks with strong ($\mu \leq 0.20$), medium ($0.20 < \mu < 0.40$), and weak ($\mu \geq 0.40$) community structures. We consider communities uncovered by Infomap as our reference case. For brevity, we report one network of each category for all the centrality measures under study in Fig 4.7.

The community-aware ranking scheme outperforms the classical descending order ranking scheme in networks with a strong community structure ($\mu \leq 0.20$). The distinction lies in the gain in the relative difference of the outbreak size (ΔR). As depicted by the U.S. Airports network (with $\mu = 0.08$) in Fig 4.7, one can note the outperformance of Closeness centrality, with a difference in the outbreak size (ΔR) reaching a maximum of 21% when the fraction of initially infected nodes (f_o) amounts to 0.41, followed by Degree centrality with ΔR = 20% at f_o = 0.41 and MNC and Katz with ΔR = 20% at f_o = 0.40. Then comes PageRank with ΔR = 15% at f_o = 0.41 followed by Betweenness with ΔR = 14% at f_o = 0.41. In general, in all the networks, Closeness, Degree, MNC, and Katz show higher ΔR than Betweenness and PageRank. One can also note three typical behaviors for the performance of the community-aware ranking scheme in networks with a strong community structure. These behaviors are common to all the centrality measures within a given network. For brevity, we report the results of Degree centrality only. The first typical behavior is that ΔR increases as f_o increases. It is illustrated by the Princeton network in Fig 4.8A

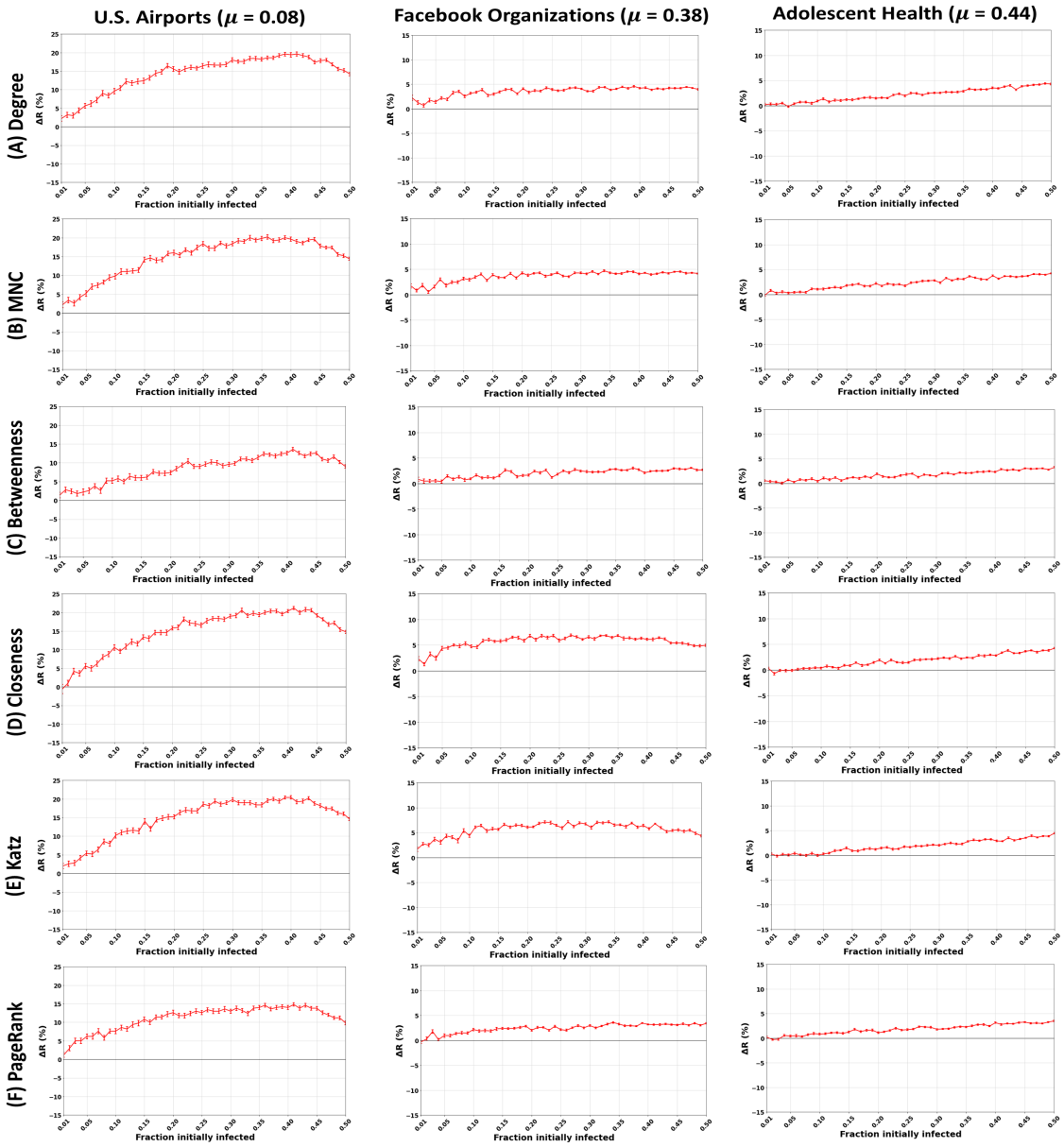


Figure 4.7: **Impact of the community structure strength (μ) in real-world networks.** The figures represent the relative difference of the outbreak size (ΔR) as a function of the fraction of initially infected nodes. The red curve indicates the relative performance difference of the community-aware ranking strategy with the descending order ranking for the six centrality measures under test. A strong, medium, and weak mixing parameter (μ) is derived based on the communities in real-world networks (U.S. Airports, Facebook Organizations, and Adolescent Health) identified by the Infomap community detection algorithm.

on the left. Ego Facebook, Facebook Friends, and Facebook Politician Pages share similar behavior. The second typical behavior is that ΔR increases until it reaches a plateau or barely deviates as f_o increases. It is shown in the middle of Fig 4.8A for the Yeast Collins network. London Transport, Malaria Genes, NetSci, Board of Directors, and DNC Emails show similar behavior. Finally, in the third case, ΔR increases until it reaches a specific

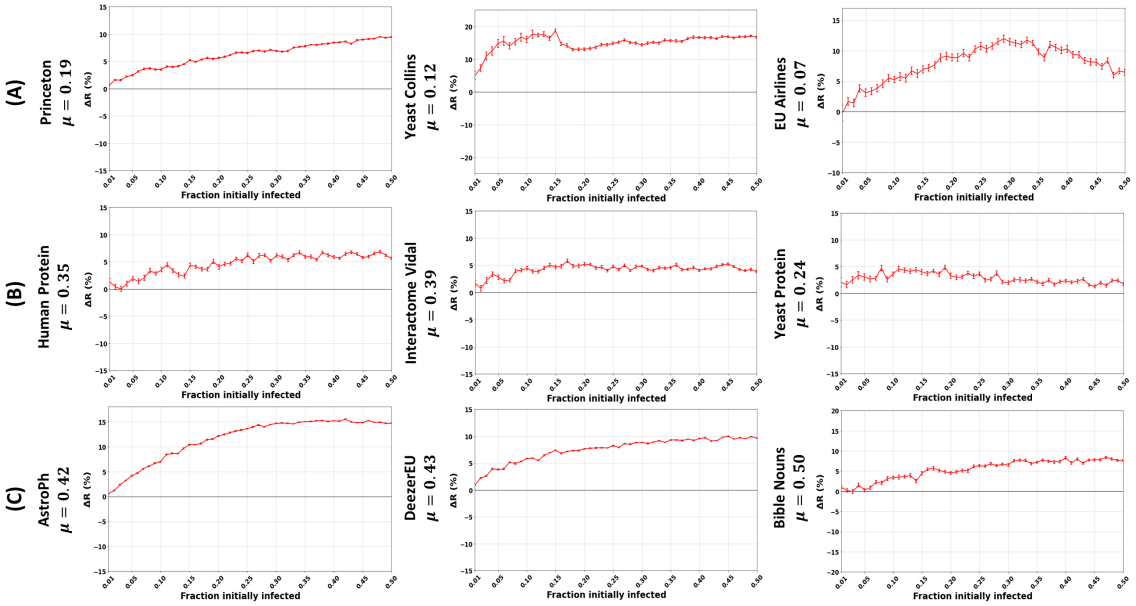


Figure 4.8: **Trends in the performance of the community-aware ranking scheme in real-world networks.** The figures represent the relative difference of the outbreak size (ΔR) as a function of the fraction of initially infected nodes. The red curve indicates the relative performance difference of the community-aware ranks of the Degree centrality compared to the descending order ranks. Communities are identified using Infomap. (A) Networks with a strong community structure strength. (B) Networks with a medium community structure strength. (C) Networks with a weak community structure strength.

value of f_o, then decreases gradually. It is demonstrated by the EU Airlines network in Fig 4.8A on the right. U.S. Airports, Madrid Train Bombings, Reptiles, 911 All Words, Marvel Partnerships, U.S. Power Grid, PGP, EuroRoad, and Internet Topology Cogentco share similar behavior.

The community-aware ranking scheme still outperforms the classical descending order ranking scheme in networks with a medium community structure ($0.20 < \mu < 0.40$). However, the gain is less pronounced compared to networks with a strong community structure. It is depicted by the Facebook Organizations network in Fig 4.7. One can see that Closeness and Katz centrality measures achieve the highest gain in ΔR amounting to 6% and 7% from a fraction of initially infected nodes amounting to 0.15 to 0.45. The maximum gain for Degree and MNC is 4% and 4.5%, respectively. Then, Betweenness and PageRank show a gain of ΔR = 2.5% and ΔR = 3.5, respectively. One can also note that within this category, we observe three behaviors for the performance of the community-aware ranking scheme. These behaviors are similar to those in networks with a strong community structure but at a smaller magnitude. We have an increasing ΔR as f_o increases, depicted by the Human Protein network on the left of Fig 4.8B. Hamsterster and Blumenau Drug share similar behavior. In the second category, we have an increasing ΔR until it reaches a plateau or barely deviates as f_o increases. It is depicted by the Interactome Vidal network in the middle of Fig 4.8B. Facebook Organizations and Caltech share similar behavior. Finally, the third category shows a slight and gradual decrease directly from the start as f_o increases. It is illustrated by the Yeast Protein network in Fig 4.8B on the right. Retweets Copenhagen shows similar behavior.

The community-aware ranking scheme outperforms the classical descending order ranking scheme in networks with a weak community structure ($\mu \geq 0.40$). However, in some networks, the gain in ΔR can even be higher than in networks with a strong or medium community structure. Indeed, the maximum improvement in ΔR can reach up to 15% in the AstroPh network (see Fig 4.8C) with Degree, MNC, Closeness, and Katz centrality measures and up to 11% and 10% for PageRank and Betweenness centrality measures respectively. At the same time, it can be as low as 5% in Adolescent Health given in Fig 4.7 for Degree, MNC, Closeness, and Katz and as low as 4% for Betweenness and 3% for PageRank. That being said, in networks with a weak community structure, there is one trend despite the difference in magnitude. Indeed, as seen from Fig 4.8C, all networks (AstroPh, DeezerEU, and Bible Nouns) have an increasing ΔR as f_o increases. The only difference is in the magnitude of ΔR from one network to another.

In summary, the community-aware ranking scheme outperforms the descending order ranking scheme in all real-world networks under study. The gain of the proposed ranking scheme is affected by the community structure strength, as observed in artificial networks with controlled community structure strength. The stronger the community structure, the higher the performance of the community-aware ranking scheme. Nevertheless, it is worth noting that the community-aware ranking also shows high performance in some real-world networks with a weak community structure.

The community-aware ranking scheme has a high performance in networks with strong community structure strength because it does not select nodes in one dense region when there are many well-separated dense areas. We visualize two networks with strong community structure strength but different topological structures, namely Yeast Collins and EU Airlines in Fig 4.9A and Fig 4.9B, respectively. In these two networks, we pick and increase the size of the top 15% of nodes selected by the descending order ranking scheme and the community-aware ranking scheme. For brevity, we only show Degree centrality and Closeness centrality. Concerning the Yeast Collins network, for both Degree and Closeness centrality, one can directly point out how the descending order scheme selects most of the top nodes in large network communities mainly located at the bottom of the network. On the contrary, the community-aware ranking scheme selects nodes in every community, spreading across all the network regions. A similar interpretation goes for the EU Airlines network, another network with a strong community structure. Indeed, the descending order ranking scheme of Degree and Closeness centrality measures targets only the dark pink and green communities. In contrast, the community-aware ranking scheme does not miss a single community. It is the reason why the community-aware ranking strategy allows a higher diffusion.

We also investigate the case of networks with a weak community structure where the proposed community-aware ranking scheme performs well. Fig 4.9C visualizes the AstroPh network, a network with a weak community structure and high performance of the community-aware ranking scheme. Despite having loosely defined communities with a vast number of inter-community connections, the community-aware ranking scheme targets nodes at the core and in the periphery of the AstroPh network, either with Degree or Closeness centrality measures. While with the descending order ranking scheme, using Degree or Closeness centrality measures, nodes picked are mainly in the network's core. Consequently, the community-aware ranking scheme can ignite a higher diffusion as it reaches regions the descending order ranking scheme never targets.

It is worth mentioning that networks characterized by a weak community structure may

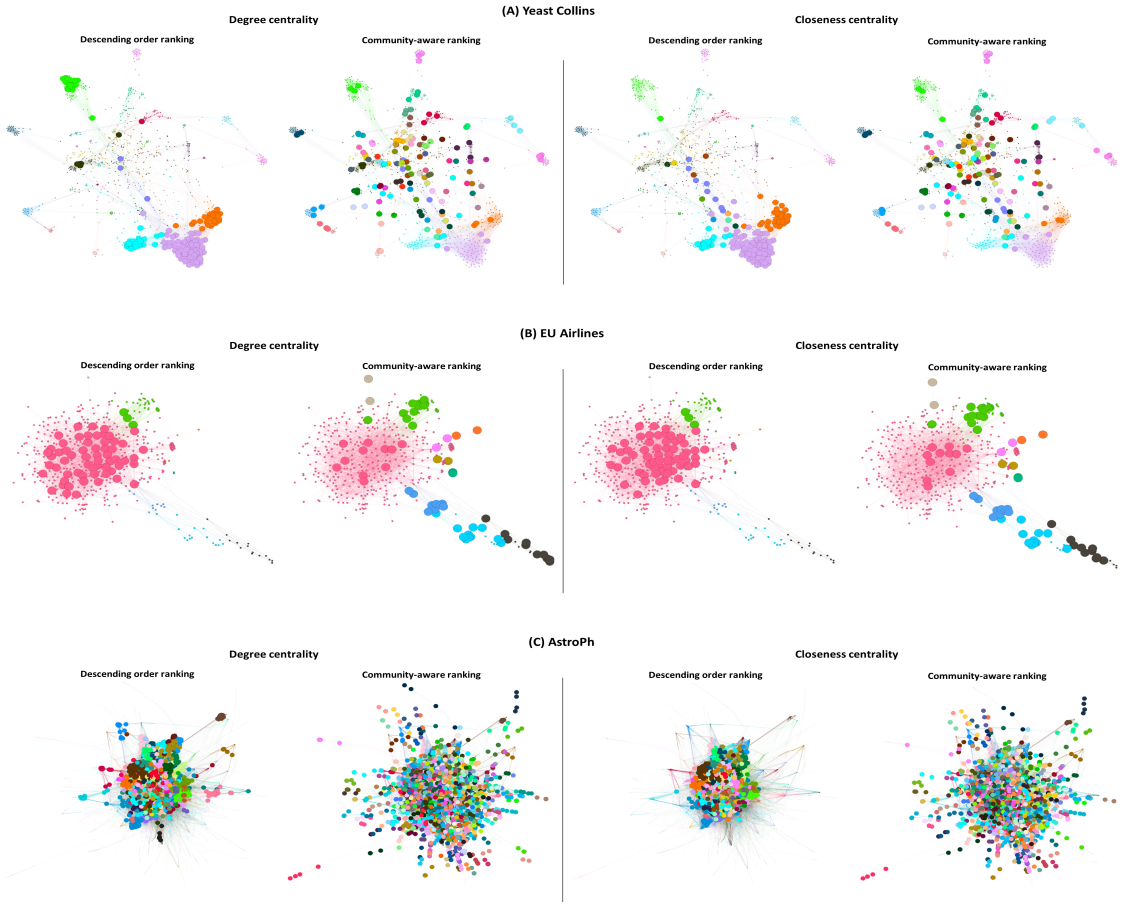


Figure 4.9: **Top nodes selected based on the Degree and Closeness centrality measures according to the descending order and community-aware ranking schemes.** Communities of the Yeast Collins (A), EU Airlines (B), and AstroPh (C) networks are identified by Infomap. The top selected nodes (depicted in bigger sizes) amount to 15% of the network's size.

exhibit different topologies [122, 123]. If it is core-periphery-like, such as in the AstroPh network, the community-aware ranking scheme covers all the regions in the network. If the network is very dense with no particular local structure, the community-aware ranking scheme might select nodes in the vicinity of each other. We suggest using a measure that combines the local and global influence of the nodes for better targeting influential nodes [117, 124]. Note that the distance between the nodes should also be addressed. Alternatively, one can also incorporate a minimum distance constraint between nodes in a community so that targeted nodes are scattered. There is room for improvement in networks with a weak community structure.

Influence of the Community Detection Algorithm

In this experiment, we use the Infomap [106] community detection algorithm to extract the communities in real-world networks. Then, we perform the same comparative evaluation process using SIR simulations between the classical and the proposed ranking strategies based on the communities identified by Louvain [87]. The aim is to investigate the

impact of the variations in the community structure induced by the community detection algorithms on the performance of the ranking schemes.

Fig 4.10 illustrates the relative difference in the outbreak size (ΔR) of the community-aware ranking scheme for Degree and Closeness centrality measures. We comment on the results for three typical networks (Facebook Friends, Human Protein, and Bible Nouns). Facebook Friends and Human Protein belong respectively to the strong and medium community structure categories using Infomap or Louvain. Bible Nouns network is in the medium community structure category based on Louvain and the weak community structure using Infomap.

First, these results demonstrate that the community-aware ranking scheme is robust to the community structure variation induced by the community detection algorithm. Indeed, ΔR is positive whether Infomap or Louvain uncovers communities. This result is independent of the community structure strength. For example, Facebook Friends is a network with a strong community structure. For a fraction of initially infected nodes equal to 0.25, the gain of ΔR equals 15% for Degree centrality and 11% for Closeness centrality. It compares to an 8% increase for Degree centrality and 7% for Closeness centrality using Louvain with the same fraction of initially infected nodes. Consider the Human Protein network with a medium community structure strength. With a fraction of initially infected nodes equal to 0.25, the ΔR gain is 6% for Degree and Closeness centrality measures. In the same situation, using Louvain, the growth is lower. Indeed, ΔR equals 2% for Degree centrality and 3% for Closeness centrality. Finally, the community-aware ranking scheme still outperforms the classical descending order ranking scheme in the Bible Nouns net-

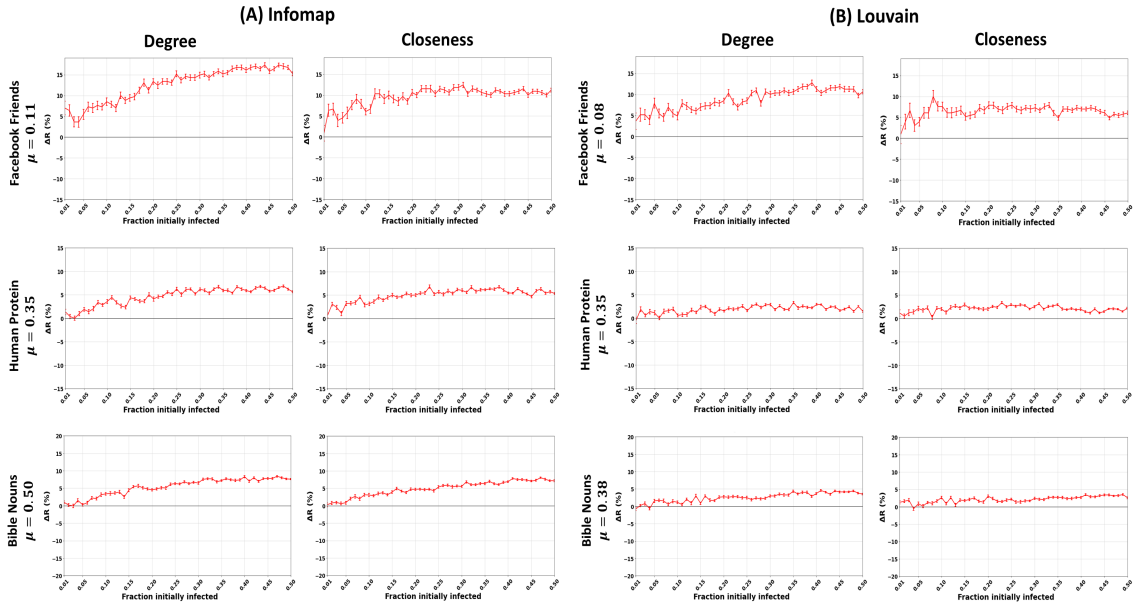


Figure 4.10: **Impact of the community detection algorithm in real-world networks with strong, medium, and weak community structure strengths.** The figures represent the relative difference of the outbreak size (ΔR) as a function of the fraction of initially infected nodes. The red curve indicates the relative performance difference of the community-aware ranks of the Degree and Closeness centrality measures compared to the descending order ranks. (A) Communities identified using Infomap. (B) Communities identified using Louvain.

work. For a fraction of initially infected nodes of 0.25, ΔR equals 6% for Degree centrality and 5.5% for Closeness centrality. Using Louvain with the same fraction of initially infected nodes reduces the ΔR gain to 2.5% for Degree centrality and 2% for Closeness centrality.

Second, one can note that with Infomap, the gain in ΔR is relatively higher than Louvain. For instance, in Facebook Friends with communities identified using Infomap, the maximum ΔR reaches 17.5% at a fraction of initially infected nodes of 0.44 using Degree centrality and 13% using Closeness at a fraction of initially infected nodes of 0.31. In contrast, with the Louvain community detection algorithm, the maximum ΔR reaches 12.5% at a fraction of initially infected nodes of 0.39 using Degree centrality and 10% using Closeness at a fraction of initially infected nodes of 0.08. Thus, the difference in gain of ΔR is +5% for Degree centrality and +3% for Closeness centrality using Infomap. One observes similar results for Human Protein and Bible Nouns. The maximum gain in ΔR in Human Protein amounts to 6% using both Degree and Closeness centrality measures. Meanwhile, Louvain's maximum gain in ΔR amounts to 3% for the two centrality measures. In Bible Nouns, with Infomap, the maximum improvement in ΔR amounts to 8.5% using both Degree and Closeness centrality measures. In opposition, using Louvain, the maximum gain in ΔR is 4.8% for Degree centrality and 3.8% for Closeness centrality.

To investigate why the performance of the community-aware ranking scheme decreases with Louvain compared to Infomap, we examine the community size distributions of the networks. Fig 4.10 gives the community size distributions associated with Infomap and Louvain of the networks provided in Fig 4.11. Indeed, they represent typical cases. Simultaneously, we compute the number of communities uncovered by each community detection algorithm and the minimum and maximum size of the communities for all the networks. Tables 5 and 6 in Appendix I report the results using Infomap and Louvain, respectively.

The histograms of the community size distributions in Fig 4.11 show that they are generally more skewed to the right using Infomap. A high number of small communities coexists with very few large communities. In contrast, this distribution is more uniform for the Louvain community structure. We observe many communities with medium and large sizes. Moreover, Infomap uncovers a higher number of communities. For example, it reveals 21 communities in Facebook Friends, 99 in Human Protein, and 88 in Bible Nouns. It compares with 10 communities in Facebook Friends, 14 in Human Protein, and 17 in Bible Nouns identified by Louvain.

The question is, how do these two outcomes affect the performance of the proposed community-aware ranking scheme compared to the classical descending order ranking scheme. On the one hand, Louvain generally uncovers fewer communities, with many medium-sized and large communities making up most of the network. On the other hand, Infomap discovers many more communities with a high number of small communities coexisting with fewer larger ones. The proposed community-aware ranking strategy is more effective with Infomap. Naturally, with more communities, it selects at least one node in every community. The higher the number of communities, the higher its ability to choose distinct nodes. It is not necessarily true with fewer communities. Indeed, first, one selects a node in each community. Then, in the next iteration, nodes are targeted in the same communities. Thus, nodes inside the same communities may be closer to each other than when there are many communities. Thus, the diffusion power weakens. Indeed, it does not reach distant regions in the network, and the diffusion stays confined in large

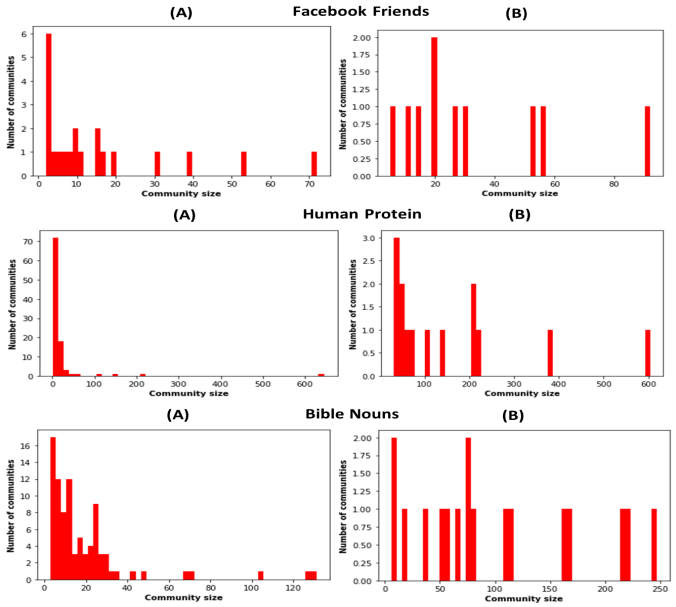


Figure 4.11: **Histograms of the community size distribution.** Communities are identified in Facebook Friends, Human Protein, and Bible Nouns by the Infomap (A) and Louvain (B) community detection algorithms.

communities. Note that these results complement those reported using the generated LFR networks with controlled community size distribution exponent regarding the number of communities. Indeed, when fewer communities exist, the community-aware ranking scheme is more susceptible to picking the next top node close to the former top-picked node inside these communities. Consequently, infecting the most influential nodes in fewer communities is less efficient than having many communities spanning the whole network.

4.5/ CONCLUSION

This chapter presents a community-aware ranking scheme that one can use with any centrality measure. The proposed method is simple yet effective at selecting nodes according to their relative influence in a modular network. Unlike the popular descending order ranking scheme, which ranks the most influential nodes from high to low centrality values, it ranks nodes in a sequential order linked to the community size. Consequently, it selects nodes across all regions of the networks. In contrast with the descending order ranking scheme that can select nodes in a few communities that ignore large parts of the network, this strategy targets nodes more uniformly distributed. As a result, the proposed strategy warrants that the diffusion process does not die out locally and reaches distant regions of the network.

Extensive experiments have been conducted on synthetic and real-world networks using the SIR epidemic spreading model. To better understand the interplay between the community structure and the performance of the proposed strategy, we performed a series of experiments in synthetic networks controlling the community structure strength, the exponents of the community size, and degree distributions. Results show that the

community-aware ranking scheme is more effective in networks with strong community structure strength. As it weakens, the performance decreases gradually. Nevertheless, the community-aware ranking scheme always performs better than the descending order ranking scheme, even in networks with a weak community structure. The community size distribution also affects the performance of the community-aware ranking scheme. Results show the strategy performs better in networks of many small communities instead of a few large communities. Indeed, the higher the number of communities, the more likely the targeted nodes are scattered across the network regions, igniting a higher epidemic outbreak. The influence of the degree distribution exponent is less pronounced. However, one can notice that the community-aware ranking scheme performs better in hub-and-spoke-like networks than in random-like ones.

The community-aware ranking scheme also outperforms the classical ranking strategy in a set of real-world networks of various domains. The findings are consistent with the synthetic networks' experiments. Indeed, the community-aware ranking scheme performs better in networks with a strong community structure strength. The gain gradually decreases with the community structure strength. Note that in some networks with a weak community structure strength, the community-aware ranking scheme still creates a higher outbreak than its alternative. Indeed, the community-aware ranking scheme ranks the top nodes inside each community. Even in networks with a weak community structure, it can rank nodes in faraway regions, causing a higher outbreak. We also investigate the influence of the community detection algorithm on the performance of the community-aware ranking scheme. The comparisons involve Infomap and Louvain. Since the community structure uncovered by Louvain results in fewer communities and subsequently larger ones compared to Infomap, the community-aware ranking scheme performs better with the Infomap community structure. This result is coherent with synthetic networks' community size distribution variation. Whatever the community detection algorithm, the community-aware ranking scheme consistently outperforms the descending order ranking strategy.

The main lesson from this study is to highlight the necessity of incorporating the community structure information in centrality measurements to better rank influential nodes. This work departs from previous community-aware solutions combining a node's local and global influence. Here, we show that whatever the notion of influence, the ranking strategy is a critical factor in the diffusion process. Whatever the centrality measure, the proposed ranking scheme is decisive in targeting the most influential nodes scattered across the network. This strategy overcomes a frequent drawback in centrality measures using the popular descending order ranking scheme in which the most influential nodes happen to be in the vicinity of each other. The proposed ranking scheme is adequate for igniting higher diffusion for marketing and awareness campaigns or combating diseases and unwanted viruses since it pinpoints influential nodes while ensuring that all regions in the network are covered.

5

The Overlapping Modularity Vitality Framework

Contents

5.1/ Introduction

Real-world networks often exhibit a community structure [125]. The recent trend of research on centrality measures exploits network community structure to design the so-called "community-aware" centrality measures [20–27]. These measures assume that a node belongs to one and only one community. While in many situations, nodes may belong to several communities, indicating an overlapping community structure [126]. We, as humans, belong to several communities merging our personal and professional lives. A protein may belong to different protein complexes. Researchers have shown that the overlap is a significant feature in many real-world networks [126–128].

While there are plenty of works to detect overlapping communities [129–135], few scientists make use of the overlapping community structure to identify critical nodes [28–31]. This chapter aims to develop a flexible framework to identify influential nodes in networks with an overlapping community structure. It allows exploiting the structure of overlapping communities pervasive in many real-world networks through various definitions of overlapping modularity. The proposed approach can be modified to suit various situations and data availability by including various types of overlap information. This allows researchers and practitioners to tailor the framework to various situations depending on the available information.

The proposed framework is inspired by the concept of vitality [136]. Given a quality function computed on the graph, the vitality of a node expresses the quality function variation when one removes it from the network. Depending on the node's role in the given

quality function, it is either negative or positive. Originally, vitality was used with classical centrality measures. Recently, [26] introduced "Modularity Vitality." It uses Newman's non-overlapping modularity as a quality measure to quantify the centrality of nodes.

When it comes to community structure, numerous quality metrics have been proposed in the literature [81, 82]. Since overlapping communities are a significant characteristic of real-world networks, we opt for overlapping information to compute the vitalities of nodes. We also rely on modularity as it is the most influential quality measure. The proposed measure, called Overlapping Modularity Vitality, is based on a generalized modularity equation that accounts for the overlapping community structure in a network. One can distinguish fuzzy and crisp overlapping modularity measures [83, 92, 137]. The proposed framework differs from its non-overlapping counterpart in its flexibility. We do not focus on a unique definition of overlapping modularity. We propose a general framework that allows including different types of overlapping structural information of the network. Indeed, one cannot extract the same information from all networks. Accordingly, depending on availability, the proposed framework can work with fuzzy and crisp information.

In the following sections, we introduce the Overlapping Modularity Vitality framework in which one can integrate multiple definitions of overlapping modularity via different formulations of the community membership strength of the nodes. We then investigate three overlapping modularity alternative definitions that incorporate contextual information about the nodes in real-world networks. The proposed framework under the three definitions is compared to its non-overlapping version and state-of-the-art overlapping centrality measures under an SIR spreading scenario. We also evaluate the influence of the community structure variation induced by the community detection algorithm on the frameworks' performance.

The chapter's main findings summarize as follows:

1. With limited resources, Overlapping Modularity Vitality with reciprocity membership targeting hubs first results in the best performance alongside OverlapNeighborhood, which targets random neighbors of overlapping nodes.
2. When resources are available, Overlapping Modularity Vitality with node similarity and degree membership to quantify a node's community membership strengths perform better when the ranking scheme is based on targeting bridges first.
3. Results demonstrate the superiority of Overlapping Modularity Vitality over its alternatives. Furthermore, one can tailor the measure to the budget using an appropriate definition of the community membership strengths and ranking strategies.

5.2/ STATE OF THE ART

This section presents the main centrality measures considering the overlapping nature of communities. In [28], L. Hébert-Dufresne *et al.* proposed an overlapping measure called "Membership." It orders nodes for immunization based on their number of shared communities. It is a local measure since it doesn't require complete network knowledge. Hence, it is suitable for large-scale networks, and it is not too sensitive to incorrect or incomplete data. Overlapping nodes with a high membership facilitate the epidemic spreading between communities. The measure has been evaluated using the SIS and SIR epidemic

models with a set of real-world networks (social, technological, and communication networks). Results show that when the disease is highly infectious, and the communities are dense, Membership outperforms coreness, degree, and betweenness centrality.

Kumar *et al*. [29] proposed the OverlapNeighborhood strategy. It randomly selects neighbors of overlapping nodes as the most crucial for immunization. It is a random local method that requires only information about the overlapping nodes. Experiments show that it performs as good as betweenness centrality while requiring less information about the network. Note that OverlapNeighorhood does not use a ranking strategy as nodes are selected randomly.

Random Walk Overlap Selection (RWOS) proposed by F. Taghavian *et al*. [30] aims to detect high-degree overlapping nodes based on a random walk. A random walk starts from a node randomly selected. A list of known or extracted overlapping nodes is checked with every step the random walker takes. The node is immunized if the visited node is in the list of overlapping nodes. Else, the random walk continues its search. The process stops when one reaches the desired fraction of nodes to immunize. This method is local because, at each step, one searches only in the neighborhood of the currently visited node. Ultimately, it targets high-degree nodes as the random walk is more likely to reach such nodes. Simulation results using the SIR epidemic model on synthetic and real-world networks show that this measure performs better in networks with strong community structure, when community membership of nodes is high, and when community sizes are small.

In [31], Ghalmane *et al*. developed Overlapping Modular Centrality, a measure that quantifies the local and global centrality of the nodes using a two-dimensional vector. After choosing a classical centrality measure, the local influence depends on whether a node is an overlapping node. One computes the local centrality of overlapping nodes considering all their communities as a single community, whereas locality applies to a single community for non-overlapping nodes. The global influence is based on the global network, constructed by the inter-community links of the node.

In recent work, Magelinski *et al*. [26] proposed a non-overlapping community-aware centrality measure called Modularity Vitality. Leveraging the concept of vitality, it can target hubs and bridges based on their contribution to Newman's modularity. Removing bridge nodes increases modularity because communities become less connected. In contrast, eliminating hubs decreases modularity because communities are less dense. Therefore, the modularity variation when removing a node indicates the type and the strength of its impact on the network's cohesiveness.

All these methods present some limitations. The measures exploiting the communities' overlap use partial information about its structure. Membership is quite basic, with many ties. It cannot distinguish between highly and poorly connected overlapping nodes. OverlapNeighborhood targets hubs in the vicinity of the overlapping nodes independently of their relations. Random Walk Overlap Selection targets highly connected overlapping nodes. Finally, Overlapping Modular Centrality merges the communities of the overlapping nodes into a single community to compute the local importance, reducing the precious information of the overlap structure. These measures fail to consider how the node is embedded in its overlapping communities. In contrast, Modularity Vitality better handles the topology, but it discards the precious overlapping information. To overcome these drawbacks, we show the potential gain of using overlapping information to identify influential nodes better [32]. This work proposes a framework called "Overlapping Modularity

Vitality," which can use any fuzzy or crisp version of the overlapping modularity quality function.

The proposed framework is based on a general definition of modularity that one can tailor to the various types of information available based on the network's structure. It generalizes the Modularity Vitality of [26] to networks with overlapping communities spanning from crisp to fuzzy structures. This work aligns with studies generalizing the traditional modularity in non-overlapping community structure [86] to the more realistic overlapping community structure. It spans from centrality measures [31, 138] to network characterization measures [92] and community detection algorithms [93, 139, 140]. The main advantage of the proposed work, falling into developing centrality measures for overlapping communities, is its flexibility in incorporating various types of information from the network rather than just a single kind of information. For instance, in collaboration networks, community membership strengths could be fuzzy, indicating the percentage of affiliation of each person to different research groups [137]. In contrast, in biological networks, a protein might participate with equal importance in various modules [141]. All these cases can be incorporated into the proposed framework, indicating its flexibility in the type of information available to identify critical nodes.

5.3/ THE OVERLAPPING MODULARITY VITALITY FRAMEWORK

In this section, we present the Overlapping Modularity Vitality framework. We introduce the vitality concept. Then, we derive an efficient technique to integrate the various versions of the overlapping modularity in this framework.

VITALITY

One way to compute the importance of a node or an edge x in a graph is through the vitality measurement [136]. The vitality of a node or an edge x denotes the variation of a real-valued arbitrary quality function f computed on a graph G when one removes x from the network. More formally, a vitality index $\mathcal{V}(G, f, x)$ is equal to the difference between the value of f on G and of f on G without the element x.

$$\mathcal{V}(G, f, x) = f(G) - f(G \setminus \{x\}) \tag{5.1}$$

OVERLAPPING MODULARITY VITALITY

We adopt the overlapping modularity Q' as the quality function f. It presents several advantages. First, many popular community detection algorithms use it as an objective function. Second, it has many extensions for networks with an overlapping community structure, whether crisp or fuzzy [83, 92, 137]. Finally, overlapping modularity is a flexible quality measure that incorporates various information on the community structure through the community membership strength parameters. For example, one may think that a node located at the overlap is more influential than a node embedded in its community. Moreover, two nodes located in the overlap might have different influences depending on the number of communities they share. Unlike the basic alternative measures, overlapping

modularity can account for these types of information. Combining the adopted definition of overlapping modularity and the vitality concept allows, therefore, to propose a general framework using tailored information about the community structure to identify influential nodes. The Overlapping Modularity Vitality of a node i is consequently defined as follows:

$$\beta_{OMV}(i) = Q'(G) - Q'(G \setminus \{i\}) \tag{5.2}$$

It is important to note that the sign of the centrality indicates the node's role. A node with a positive centrality value plays a hub role. Its removal decreases the overlapping modularity of the network. Subsequently, it has a positive centrality value. A hubs-first ranking strategy indicates that one ranks nodes in descending order starting from the highest positive centrality values. In contrast, if a node acts as a bridge between two communities, its removal increases the network's modularity. Indeed, the number of inter-community links between communities decreases, leading to better-defined communities. Consequently, it receives a negative centrality value. In a bridges-first ranking strategy, one orders the nodes from the highest negative to the highest positive centrality values. To aggregate the influence of hubs and bridges simultaneously, we consider the absolute value of the centrality of each node. Then, one sorts the nodes in descending order. These ranking strategies are decisive in diffusion scenarios since the budget is limited.

Computing the Overlapping Modularity Vitality

Computing the overlapping modularity naïvely can be prohibitive for large networks. Indeed, computing the modularity is in $O(|E|)$. Therefore, computing the contribution of each node directly requires $O(N|E|)$. One needs an efficient way to update the modularity after removing a node. [26] proposed an efficient iterative process to compute the non-overlapping Modularity Vitality. Inspired by their approach, we extend their development to the Overlapping Modularity Vitality framework, whether the community membership strength of the node is crisp or fuzzy. Following is an expression to compute the modularity after the removal of node i:

$$Q'(G \setminus \{i\}) = \sum_{c_q \in C} \left[\frac{|E^{in}_{c_q}| - |E^{in}_{i,c_q}|}{|E| - |E_i|} - \left(\frac{2\left(|E^{in}_{c_q}| - |E^{in}_{i,c_q}|\right) + \left(|E^{out}_{c_q}| - |E^{out}_{i,c_q}|\right)}{2(|E| - |E_i|)} \right)^2 \right] \tag{5.3}$$

One uses $|E|$, $|E^{in}_{c_q}|$, and $|E^{out}_{c_q}|$ to calculate $Q'(G)$. The only requirement to compute the variation is to calculate the node's contribution inside and outside its communities ($|E^{in}_{i,c_q}|$ and $|E^{out}_{i,c_q}|$). The node's total contribution ($|E_i|$) is the sum over the communities of the node's internal contribution ($|E^{in}_{i,c_q}|$) and external contribution ($|E^{out}_{i,c_q}|$). Thus, the extra computation has a $O(N|C| + |E|)$ complexity for the increment.

PROPOSED FRAMEWORK

One can use the developed framework with any definition of community membership strengths. In the experiments, we investigate three versions of community membership strengths (i.e., reciprocity membership, degree membership, and node similarity). Note that we do not consider regrouping the community structure when removing a node since there is no closed-form solution for the re-computation of the overlapping modularity alongside the community membership strengths. Algorithm 2 reports the steps of the Overlapping Modularity Vitality computation. The adjacency matrix $A_{i,j}$ weighted by the crisp and fuzzy community membership strengths of the nodes is given in Algorithm 3 and Algorithm 4.

The time complexity of Algorithm 2 is $O(N|C|)$. Indeed, lines 1 to 5 require $O(|C|)$. For lines 6 till 12, it requires $O(N|C|)$. Finally, lines 13 and 14, require $O(N)$. Algorithms 3 and 4 have $O(N|C|)$ complexity. The complexity of acquiring the community membership strengths depends on the acquisition method. In our study, the complexities for reciprocity membership, degree membership, and node similarity are $O(N)$, $O(N < k >)$, and $O(2N)$, respectively.

Algorithm 2 Overlapping Modularity Vitality

Input: Graph $G(V, E)$, Weighted adjacency matrix $W_{i,j}$, Community set C
Output: Vitality of each node $\beta_{OMV}(i)$

// Computing the overlapping modularity of the graph
1: **for each** $c_q \in C$ **do**
2: Compute the total internal weights $|E^{in}_{c_q}|$ of community c_q
3: Compute the total external weights $|E^{out}_{c_q}|$ of community c_q
4: $|E| \leftarrow |E^{in}_{c_q}| + |E^{out}_{c_q}|$
5: $Q'(G) \leftarrow \sum_{c_q \in C} \left[\frac{|E^{in}_{c_q}|}{|E|} - \left(\frac{2|E^{in}_{c_q}| + |E^{out}_{c_q}|}{2|E|} \right)^2 \right]$
// Computing the overlapping modularity increment of each node
6: **for each** $i \in V$ **do**
7: **for each** $c_q \in C$ **do**
8: **if** $i \in c_q$ **then**
9: Compute the total internal weights $|E^{in}_{i,c_q}|$ of node i
10: Compute the total external weights $|E^{out}_{i,c_q}|$ of node i
11: $|E_i| \leftarrow \sum_{c_q \in C} |E^{in}_{i,c_q}| + |E^{out}_{i,c_q}|$
12: $Q'(G \setminus \{i\}) \leftarrow \sum_{c_q \in C} \left[\frac{|E^{in}_{c_q}| - |E^{in}_{i,c_q}|}{|E| - |E_i|} - \left(\frac{2\left(|E^{in}_{c_q}| - |E^{in}_{i,c_q}|\right) + \left(|E^{out}_{c_q}| - |E^{out}_{i,c_q}|\right)}{2(|E| - |E_i|)} \right)^2 \right]$
// Computing the vitality of each node
13: **for each** $i \in V$ **do**
14: $\beta_{OMV}(i) \leftarrow Q'(G) - Q'(G \setminus \{i\})$

Algorithm 3 Integrating the Average of Community Membership Strength

Input: Graph $G(V, E)$, Adjacency matrix $A_{i,j}$, Community set C, Community membership strength of the nodes $\Theta(i)$
Output: Weighted adjacency matrix $W_{i,j}$

for each $i, j \in V$ **do**
  for each $c_q \in C$ **do**
    if $i, j \in c_q$ **then**
      $W_{i,j} \leftarrow \frac{\theta_{i,c_q}+\theta_{j,c_q}}{2} A_{i,j}$
    else
      $W_{i,j} \leftarrow \frac{\theta_{i,c_q}+(1-\theta_{j,c_q})}{2} A_{i,j}$

Algorithm 4 Integrating the Product of Community Membership Strength

Input: Graph $G(V, E)$, Adjacency matrix $A_{i,j}$, Community set C, Community membership strength of the nodes $\Theta(i)$
Output: Weighted adjacency matrix $W_{i,j}$

for each $i, j \in V$ **do**
  for each $c_q \in C$ **do**
    if $i, j \in c_q$ **then**
      $W_{i,j} \leftarrow s(\Theta(i), \Theta(j)) A_{i,j}$
    else
      $W_{i,j} \leftarrow [1 - s(\Theta(i), \Theta(j))] A_{i,j}$

5.4/ COMPARING THE COMMUNITY-AWARE CENTRALITY MEASURES

This experiment allows us to compare the performance of the Modularity Vitality measures (i.e., overlapping and non-overlapping) with their alternatives on 21 networks. The alternative measures include Membership, OverlapNeighborhood, Random Walk Overlap Selection, and Overlapping Modular Centrality. Like the previous experiment, we consider three ranking strategies for the vitality-based measures: hubs-first, bridges-first, and hubs and bridges. We also evaluate the influence of the algorithm used to uncover the community structure on the performance of the measures. Note that we use the Degree centrality (η_d) as a reference to compare all the measures in this case. Note that Table 5.1 provides the abbreviations of the centrality measures used in this chapter.

Table 5.1: The abbreviations of the centrality measures used.

Symbol	Meaning
β_M	Membership
β_{ON}	OverlapNeighborhood
β_{RWOS}	Random Walk Overlap Selection
β_{OMC}	Overlapping Modular Centrality
α^{+}_{MV}	Modularity Vitality with a hubs-first ranking strategy
α^{-}_{MV}	Modularity Vitality with a bridges-first ranking strategy
$\lvert\alpha_{MV}\rvert$	Modularity Vitality with a hubs and bridges ranking strategy
β^{R+}_{OMV}	Overlapping Modularity Vitality with reciprocity membership and a hubs-first ranking strategy
β^{D+}_{OMV}	Overlapping Modularity Vitality with degree membership and a hubs-first ranking strategy
β^{S+}_{OMV}	Overlapping Modularity Vitality with node similarity and a hubs-first ranking strategy
β^{R-}_{OMV}	Overlapping Modularity Vitality with reciprocity membership and a bridges-first ranking strategy
β^{D-}_{OMV}	Overlapping Modularity Vitality with degree membership and a bridges-first ranking strategy
β^{S-}_{OMV}	Overlapping Modularity Vitality with node similarity and a bridges-first ranking strategy
$\lvert\beta^{R}_{OMV}\rvert$	Overlapping Modularity Vitality with reciprocity membership and a hubs and bridges ranking strategy
$\lvert\beta^{D}_{OMV}\rvert$	Overlapping Modularity Vitality with degree membership and a hubs and bridges ranking strategy
$\lvert\beta^{S}_{OMV}\rvert$	Overlapping Modularity Vitality with node similarity and a hubs and bridges ranking strategy
η_d	Degree centrality
R_r	Reference centrality

HUBS-FIRST RANKING STRATEGY

We observe four typical behaviors illustrated in Figure 5.1 and the first row (A) of Figure 5.2 using the hubs-first strategy.

The first case, illustrated by the DBLP, AstroPh, and DeezerEU networks in Figure 5.1, shows that Overlapping Modularity Vitality based on node similarity outperforms at the beginning its alternative definitions (i.e., reciprocity membership and degree membership) and the remaining alternative measures. Then, either reciprocity membership or degree membership takes over.

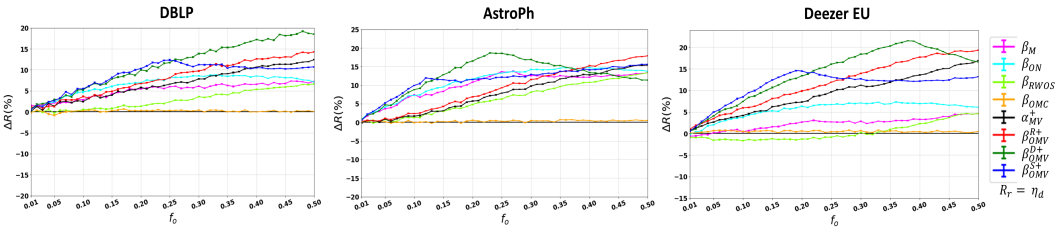


Figure 5.1: **Hubs-first ranking strategy.** The relative difference of the outbreak size (ΔR) as a function of the fraction of initially infected nodes (f_o) of the networks DBLP, AstroPh, and DeezerEU.

The second case, illustrated by Bible Nouns, demonstrates the superiority of Overlapping Modularity Vitality based on the reciprocity membership method (β^{R+}_{OMV}). There is no single winner for low values of the initial fraction of infected nodes ($f_o \leq 0.10$). However, β^{R+}_{OMV} outperforms all its alternatives when f_o increases. The gain on the reference (η_d) is around 21% for $f_o = 0.5$. It is 6% higher than the second-best performing centrality (α^+_{MV}). U.S. Airports, EU Airlines, Princeton, and 911AllWords present similar behavior.

The third case shows that when infecting a small fraction of nodes (f_o), OverlapNeighborhood (β_{ON}) performs slightly better than the other measures. Then, as the fraction of initially infected nodes increases, other centrality measures join the lead (i.e., Membership (β_M), Random Walk Overlap Selection (β_{RWOS}), Modularity Vitality (α^+_{MV}), or Overlapping Modularity Vitality with reciprocity membership (β^{R+}_{OMV})). Finally, their gain over the reference declines except for Modularity Vitality (α^+_{MV}) and Overlapping Modularity Vitality with reciprocity membership (β^{R+}_{OMV}). For example, in DNC Emails, when $f_o < 0.25$, OverlapNeighborhood shows a slightly higher relative difference in the relative epidemic outbreak size (ΔR). In the range $f_o = 0.25$ to $f_o = 0.30$, it is joined by Membership, Random Walk Overlap Selection, Modularity Vitality, and Overlapping Modularity Vitality with reciprocity membership. When $f_o > 0.30$, ΔR decreases for all the centrality measures except for Modularity Vitality and Overlapping Modularity Vitality with reciprocity membership. Facebook Politician Pages, Yeast Collins, Facebook Friends, Hamsterster, Caltech, Facebook Organizations, Budapest Connectome, Ego Facebook, and PGP show similar behavior. We note that only in Ego Facebook and PGP, Overlapping Modularity Vitality with reciprocity membership is not as effective compared to the stated well-performing centrality measures.

The fourth case, illustrated by Adolescent Health, shows that the Modularity Vitality (α^+_{MV}) is the best performing centrality starting from a small fraction of initially infected nodes (f_o). However, the difference with Overlapping Modularity Vitality based on reciprocity membership (β^{R+}_{OMV}) is pretty small at a higher fraction of initially infected nodes. Yeast Protein and Reptiles show similar behavior.

To summarize, we observe four behaviors using the hubs-first ranking strategy. First, Overlapping Modularity Vitality based on node similarity followed by degree membership

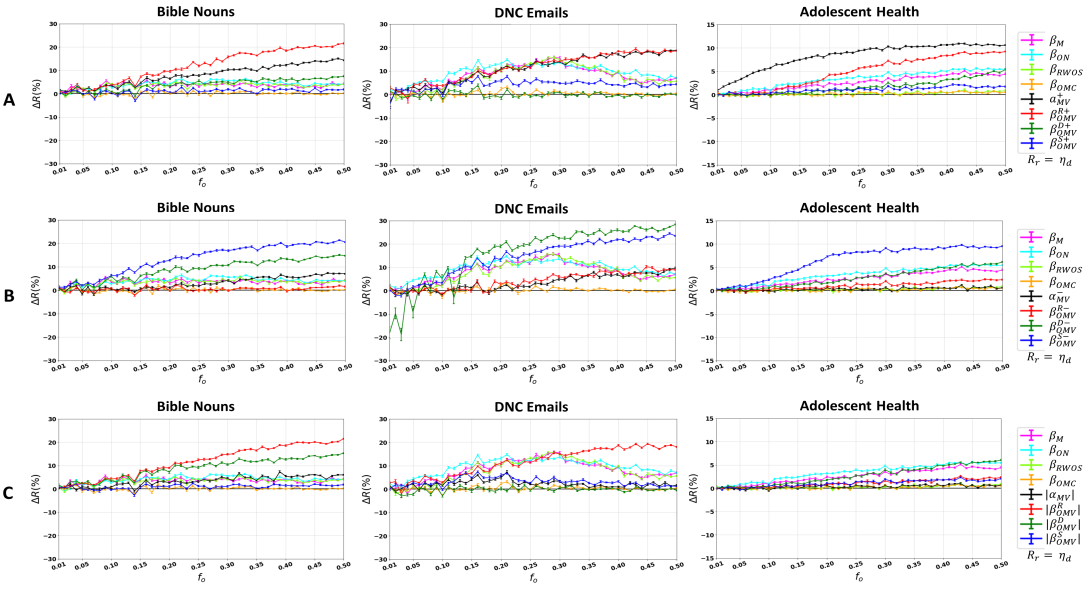


Figure 5.2: **Hubs-first (A), bridges-first (B), and hubs & bridges (C) ranking strategies.** The relative difference of the outbreak size (ΔR) as a function of the fraction of initially infected nodes (f_o) of the networks Bible Nouns, DNC Emails, and Adolescent Health.

and reciprocity membership at a higher fraction of initially infected nodes produces the highest outbreak. Second, Overlapping Modularity Vitality with reciprocity membership outperforms all its alternative measures on a group of networks. In the third case, for a small fraction of initially infected nodes, OverlapNeighborhood is the best performer before being surpassed by Modularity Vitality and Overlapping Modularity Vitality based on reciprocity membership. It suggests that one should randomly target the overlapping neighbors on a low budget to a certain extent on these networks. Then, leveraging deterministic information from the community structure is more beneficial. Finally, in the fourth case, Modularity Vitality outperforms all its alternatives. It suggests that, in this case, information about the overlap is not decisive.

Bridges-first Ranking Strategy

We still observe three typical behaviors in the 21 networks using the bridges-first ranking strategy. The results of the six networks showcasing these behaviors are given in the second row (B) of Figure 5.2 and Figure 5.3.

The first case, illustrated by Bible Nouns, shows that Overlapping Modularity Vitality based on node similarity (β_{OMV}^{S-}) takes the lead either in all the range of the fraction of initially infected nodes (f_o) or after one reaches a specific value of f_o. For example, in the Bible Nouns network, when the fraction of initially infected nodes (f_o) is less than 0.10, β_{OMV}^{S-} performs as well as Overlapping Modularity Vitality based on degree membership (β_{OMV}^{D-}). Then, it keeps increasing at a higher pace outperforming all the other measures, including β_{OMV}^{D-}. Its gain over the reference reaches 21% at $f_o = 0.49$. It is 6% more than the second-best performing centrality. Facebook Organizations, Reptiles, Ego Facebook, Facebook Politician Pages, Yeast Protein, Adolescent Health, and PGP exhibit similar behavior. We note that in the Yeast Protein network, given in Figure 5.3, the best

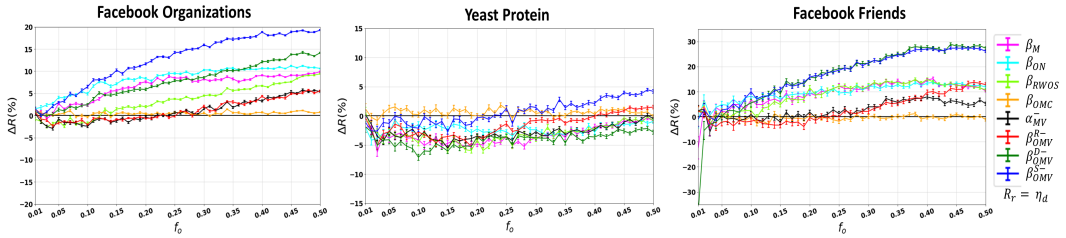


Figure 5.3: **Bridges-first ranking strategy.** The relative difference of the outbreak size (ΔR) as a function of the fraction of initially infected nodes (f_o) of the networks Facebook Organizations, Yeast Protein, and Facebook Friends.

performing centrality is first Overlapping Modular Centrality (β_{OMC}) before node similarity (β_{OMV}^{S-}) takes over. While in the other networks of this category, OverlapNeighborhood (β_{ON}) performs better when resources are limited. One can observe the initial superiority of OverlapNeighborhood (β_{ON}) in Facebook Organizations in Figure 5.3.

The second case, illustrated by DNC Emails, shows that for low values of the initial fraction of infected nodes, OverlapNeighborhood (β_{ON}) or Overlapping Modularity Vitality using node similarity (β_{OMV}^{S-}) perform the best. After a specific fraction of initially infected nodes (f_o), the outperformance is interchangeable between Overlapping Modularity Vitality based on degree membership (β_{OMV}^{D-}) and node similarity (β_{OMV}^{S-}). For example, in DNC Emails, the relative outbreak size of Overlapping Modularity Vitality based on degree membership fluctuates positively to negatively when $f_o \leq 0.14$. Above this value, it increases and outperforms the rest of the measures until $f_o = 0.50$. At this point, it exceeds the reference centrality by 29%. EU Airlines, U.S. Airports, Hamsterster, Budapest Connectome, Princeton, DBLP, AstroPh, DeezerEU, and 911AllWords show similar behavior.

Facebook Friends is a typical example of the third case. Figure 5.3 shows that two Overlapping Modularity Vitality versions (i.e., node similarity (β_{OMV}^{S-}) and degree membership (β_{OMV}^{D-})) perform similarly. They outperform the other measures after one reaches a specific fraction of initially infected nodes (f_o). Before passing f_o, OverlapNeighborhood (β_{ON}) performs better. The networks Yeast Collins and Caltech also follow a similar behavior.

Using the bridges-first strategy, we observe the effectiveness of two versions of Overlapping Modularity Vitality (i.e., node similarity and degree membership). However, the gain on the reference starts at a specific fraction of initially infected nodes. Before reaching this value, OverlapNeighborhood performs better. Thus, one should use bridges that predominantly affect the overlapping modularity when enough resources are available. With a low budget, one should target random nodes situated near overlaps of communities using OverlapNeighborhood.

Hubs and Bridges Ranking Strategy

With the hubs and bridges ranking strategy, one can distinguish three behaviors. The third row (C) of Figure 5.2 and Figure 5.4 report the results for three typical networks.

The first case, illustrated by Bible Nouns, shows that Overlapping Modularity Vitality based on the reciprocity membership method ($|\beta_{OMV}^{R}|$) ranks first after a specific frac-

tion of initially infected nodes (f_o). For instance, in the Bible Nouns network, when $f_o \geq 0.13$, $|\beta^R_{OMV}|$ begins outperforming other measures at an increasing rate. It reaches the maximum relative epidemic outbreak size (ΔR) of 21% at f_o = 0.50. Compared to the second-best performing centrality, namely, Overlapping Modularity Vitality based on the degree membership method ($|\beta^D_{OMV}|$), the difference is 6%. When $f_o < 0.13$, there is no clear winner. EU Airlines, U.S. Airports, and Hamsterster show similar behavior. The maximum relative epidemic outbreak size (ΔR) reaches 28%, 22.5%, 22%, 20%, and 12% at f_o = 0.50 in EU Airlines, Princeton, U.S. Airports, Hamsterster, and 911AllWords, respectively.

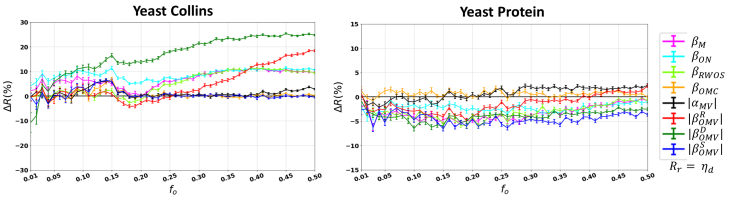


Figure 5.4: **Hubs & bridges ranking strategy.** The relative difference of the outbreak size (ΔR) as a function of the fraction of initially infected nodes (f_o) of the Yeast Collins and Yeast Protein networks.

The second case, illustrated by DNC Emails, shows that OverlapNeighborhood (β_{ON}) performs better than the other measures up to a specific fraction of initially infected nodes (f_o). Then, either Overlapping Modularity Vitality with reciprocity membership ($|\beta^R_{OMV}|$), degree membership ($|\beta^D_{OMV}|$), or node similarity ($|\beta^S_{OMV}|$) takes over. For example, in DNC Emails, OverlapNeighborhood outperforms the other measures up to $f_o = 0.22$. Then, its curve coincides with those of Random Walk Overlap Selection (β_{RWOS}), Membership (β_M), and Overlapping Modularity Vitality with reciprocity membership ($|\beta^R_{OMV}|$). At $f_o = 0.30$, all other measures decrease while $|\beta^R_{OMV}|$ keeps increasing until it reaches a gain of 19% on the reference. Similarly, Yeast Collins provided in Figure 5.4, shows that β_{ON} does slightly better up till $f_o = 0.06$. Then, degree membership ($|\beta^D_{OMV}|$) outperforms the other measures with a gain up to 25%. Ego Facebook, Facebook Politician Pages, Facebook Organizations, Facebook Friends, Caltech, Adolescent Health, Reptiles, Budapest Connectome, PGP, Yeast Collins, AstroPh, DBLP, and DeezerEU follow a similar behavior, with f_o changing from one network to another.

Yeast Protein represents the third and final case in Figure 5.4. Most overlapping centrality measures perform worse in this network than reference centrality (i.e., Degree centrality). It is not the case for Overlapping Modular Centrality (β_{OMC}) and Modularity Vitality ($|\alpha_{MV}|$). All curves show a negative relative difference in the epidemic outbreak size (ΔR) except for β_{OMC} and $|\alpha_{MV}|$. These two centrality measures outperform the reference by a maximum of 3%.

In summary, with the hubs and bridges ranking strategy and a limited budget, one should target neighbors of the overlapping nodes using OverlapNeighborhood. If enough budget is available, Overlapping Modularity Vitality based on reciprocity membership or degree membership results in a much larger epidemic outbreak than its alternatives.

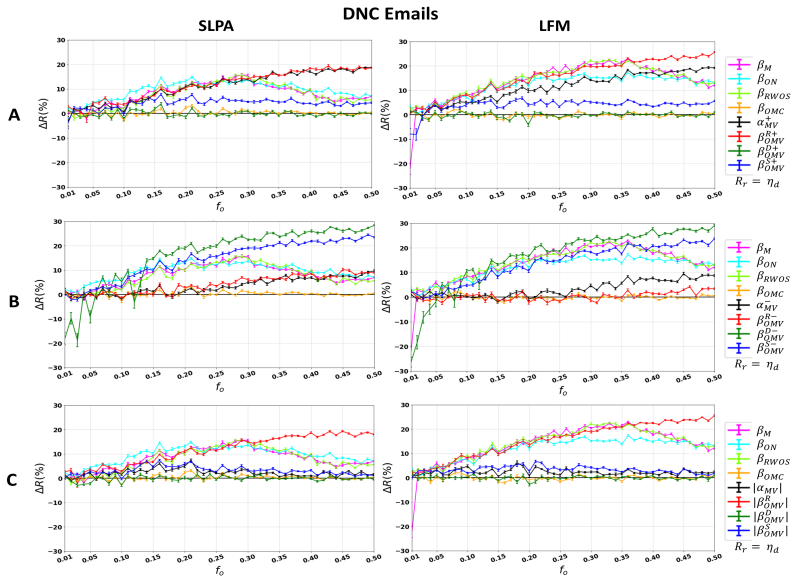


Figure 5.5: **Influence of the community detection algorithm in DNC Emails.** The relative difference of the outbreak size (ΔR) as a function of the fraction of initially infected nodes (f_o). The figures on the left represent the results of the SLPA community detection algorithm, while the figures on the right represent the results of the LFM community detection algorithm. The ranking strategies are as follows hubs-first (A), bridges-first (B), and hubs & bridges (C).

Influence of the Community Detection Algorithms

We now use the Lancichinetti Fortunato Method (LFM) to uncover the community structure of the networks. We perform the performance analysis of the centrality measures under test and compare it with the community structure results extracted with the Speaker-Listener Label Propagation Algorithm (SLPA). Our goal is to investigate the consistency of the centrality measures when one uses a different community detection algorithm. We present two extreme cases. The first case has similar communities identified by the two community detection algorithms. In contrast, the second case greatly differs in the number of communities identified.

Figure 5.5 reports the relative difference in the epidemic outbreak size in the DNC Emails network using SLPA and LFM. The community structure identified by SLPA contains 13 communities, while LFM extracts 15 communities.

Results show that the performance of the centrality measures with the three different ranking strategies (i.e., hubs-first, bridges-first, and hubs and bridges) is similar. One can see slight differences in Overlapping Modularity Vitality based on reciprocity membership. Its gain on the reference reaches 25% with the community structure uncovered by LFM at f_o = 50% with the hubs-first and hubs and bridges ranking schemes. It is to compare to a 20% gain with the community structure of SLPA. In addition, OverlapNeighborhood and Membership perform slightly better with LFM. More specifically, the gain of OverlapNeighborhood reaches 13% at f_o = 0.50 while it is only 8% with SLPA. Membership reaches its highest gain of 21% at f_o = 0.36 using LFM compared to 15% at f_o = 0.30 with SLPA.

Figure 5.6 presents the performance of the centrality measures in the Bible Nouns network using SLPA and LFM. SLPA identifies 13 communities, while LFM identifies 150

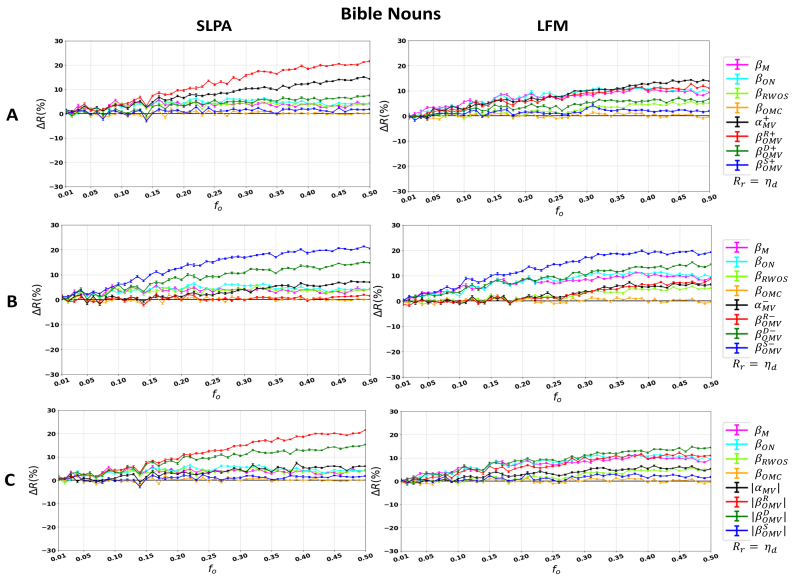


Figure 5.6: **Influence of the community detection algorithm in Bible Nouns.** The relative difference of the outbreak size (ΔR) as a function of the fraction of initially infected nodes (f_o). The figures on the left represent the results of the SLPA community detection algorithm, while the figures on the right represent the results of the LFM community detection algorithm. The ranking strategies are as follows hubs-first (A), bridges-first (B), and hubs & bridges (C).

small communities. In this situation, Overlapping Modularity Vitality based on reciprocity membership deteriorates when the ranking strategy is hubs-first or hubs and bridges simultaneously. More specifically, using the hubs-first approach and SLPA, the gain on the reference reaches 21% at f_o = 0.50 compared to 11% when LFM uncovers the community structure. One observes a similar behavior using the hubs and bridges ranking strategy. The higher the number of communities, the less effective Overlapping Modularity Vitality based on reciprocity membership is. It is because its ability to differentiate the nodes decreases when the number of communities increases. Consider two extreme cases: a network with 13 overlapping nodes belonging to 13 communities and another network with 150 overlapping nodes belonging to 150 communities. Even though all nodes have the same reciprocity membership value, the number of nodes with distinct centrality values is higher in the first case. Therefore, it is more discriminative when the number of communities is low.

To conclude, the main driver of performance sensitivity to the community structure variation is the number of communities. Indeed, overall performance evolution exhibits similar trends for the various centrality measures when the number of communities uncovered by the community detection algorithms is comparable. In contrast, if the number of communities differs significantly, Overlapping Modularity Vitality based on reciprocity membership underperforms when the number of communities is high. In this case, more overlapping nodes share the same centrality value.

5.5/ DISCUSSION

We now investigate the underlying behavior of the best performing centrality measures. We explore why Overlapping Modularity Vitality based on reciprocity membership targeting hubs performs best when resources are limited. We also examine why Overlapping Modularity Vitality targeting bridges based on degree membership and node similarity are more effective with more resources. A node is bridge-oriented if it has more inter-community links than intra-community links. In addition, a node is hub-oriented if it has more intra-community links than inter-community links. Figure 5.7 shows a toy example of four networks with increasing communities. In this example, node 3 is the only overlapping node. In subfigure A, containing two communities, node 3 is hub-oriented since all of its links are intra-community links.

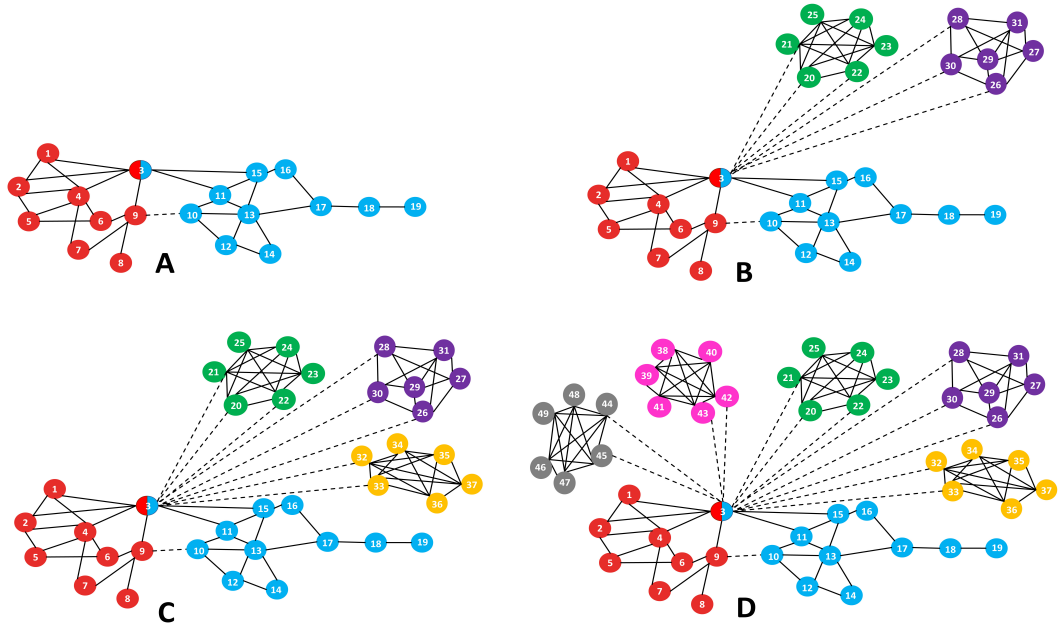


Figure 5.7: **A toy example illustrating the impact of removing an overlapping node in four different networks.** The number of communities in these networks ranges from two to seven. Node 3 is the only overlapping node. The three versions of Overlapping modularity vitality (reciprocity membership, degree membership, and node similarity) detect it as an overlapping bridge in networks B, C, and D, respectively. Solid lines represent intra-community links, and dashed lines represent inter-community links.

As the number of communities grows, node 3 transitions from an overlapping hub-oriented node to an overlapping bridge-oriented node. Reciprocity membership is the first version of Overlapping Modularity Vitality to identify node 3 as an overlapping bridge (i.e., it receives a negative value). It starts in subfigure B, where node 3 has the same number of intra-community and inter-community links. Degree membership begins by identifying node 3 as an overlapping bridge in subfigure C. It has more inter-community links than intra-community links in this case. Finally, node similarity identifies node 3 as an overlapping bridge in subfigure D. It has a higher number of inter-community links in this case. This example shows that reciprocity membership is more effective at detecting bridges. Degree membership follows, and node similarity is the last. Table 5.2 reports the

Table 5.2: Overlapping Modularity Vitality values of node 3 in the toy example using the three different approaches: reciprocity membership (β^{R}_{OMV}), degree membership (β^{D}_{OMV}), and node similarity (β^{S}_{OMV}).

Network	A	B	C	D
Reciprocity membership (β^{R}_{OMV})	0.254	**-0.070**	-0.138	-0.203
Degree membership (β^{D}_{OMV})	0.434	0.069	**-0.009**	-0.093
Node similarity (β^{S}_{OMV})	0.033	0.004	0.001	**-0.0001**

centrality values of node 3 in the four networks. If there are more overlapping nodes, then as reciprocity membership is prompt at identifying bridges, they are more likely to be near each other. In contrast, since degree membership and node similarity are less sensitive to bridges, they tend to be far apart. Consequently, those nodes distributed across all communities diffuse the epidemics in communities unreachable with clustered bridges.

In contrast, with the hubs-first ranking strategy, the effectiveness of Overlapping Modularity Vitality based on reciprocity membership is due to its ability to select hubs in communities far apart. Consequently, distinct nodes are exposed to the epidemic, causing a higher virus circulation under limited resources. However, since the distance between those nodes is not as high as the ones picked by Overlapping Modularity Vitality based on degree membership and node similarity targeting bridges, the diffusion does not scale up when more nodes are infected.

Similarly is the case with the performance of OverlapNeighborhood and Modularity Vitality targeting hubs when resources are limited. They also rely on their ability to pick up nodes far away from each other. However, their potential in picking up distant nodes does not augment as Overlapping Modularity Vitality using degree membership and node similarity targeting bridges. Since hubs are more frequent in large communities, removing a few does not significantly affect the network's modularity. It is the opposite with smaller communities. This is the reason why Modularity Vitality targeting hubs falls behind. Indeed, it targets hubs located in smaller communities first. As small communities are far apart, when resources are limited, it is more beneficial to target their hubs. However, as the availability of resources increases, one needs to target hubs in small and large communities.

We use the Yeast Collins and Facebook Friends networks to visualize the nodes targeted by the overlapping centrality measures. Figure 5.8 and Figure 5.9 show the initially infected nodes for Modularity Vitality and the three versions of Overlapping Modularity Vitality in the stated networks. They visually depict the infected nodes of the vitality measures with a low fraction of initially infected nodes ($f_o = 2\%$) and a high fraction of initially infected nodes ($f_o = 20\%$) using a hubs-first ranking strategy (indicated with a $+$ sign) and a bridges-first ranking strategy (marked with a $-$ sign). One can see that in both networks, using Overlapping Modularity Vitality based on degree membership and node similarity targeting bridges (β^{D-}_{OMV}, β^{S-}_{OMV}), the top 2% and top 20% nodes initially infected are more distant as compared to the Modularity Vitality targeting bridges (α^{-}_{MV}) and Overlapping Modularity Vitality using reciprocity membership targeting bridges (β^{R-}_{OMV}).

When resources are limited, we observe two behaviors resulting in the largest outbreak. Either infecting hubs inside different communities using Overlapping Modularity Vitality with nodes reciprocity targeting hubs first or infecting the neighbors of overlapping nodes using OverlapNeighborhood centrality. We show the nodes selected by the Overlap-

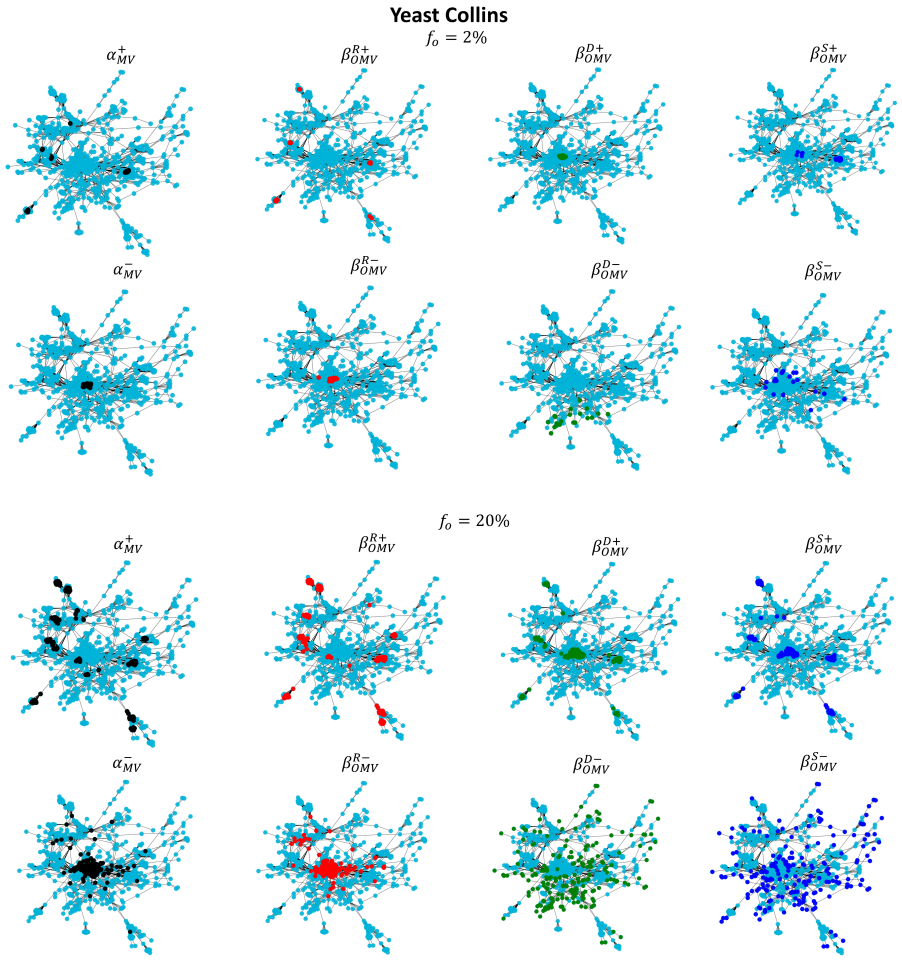


Figure 5.8: **The Yeast Collins network with the nodes chosen to be initially infected by the vitality measures**: Modularity Vitality (α_{MV}), and the three different versions of Overlapping Modularity Vitality, namely: reciprocity membership (β^{R}_{OMV}), degree membership (β^{D}_{OMV}), and node similarity (β^{S}_{OMV}). The measures use a hubs-first ranking scheme denoted by a "+" sign and a bridges-first ranking scheme marked by a "$-$" sign. The top two rows and the bottom two rows represent 2% and 20% of the fraction of initially infected nodes, respectively.

Neighborhood centrality in Figure 5.10 for the Yeast Collins and Facebook Friends networks. In Facebook Friends, when resources are low, OverlapNeighborhood centrality performs slightly better. Let's compare it with the nodes chosen by the vitality measures in Figure 5.9. At $f_o = 2\%$, nodes targeted by OverlapNeighborhood are more scattered across the network, resulting in a slightly higher outbreak. Similarly is the case for the Yeast Collins network.

On the contrary, it is better to target distant bridges when enough resources are available. Indeed, an epidemic can reach the entire network without getting stuck inside its originating community with this strategy. These results demonstrate the importance of incorporating information on the overlapping communities to ignite a more extensive epidemic outbreak capable of spreading across the whole network.

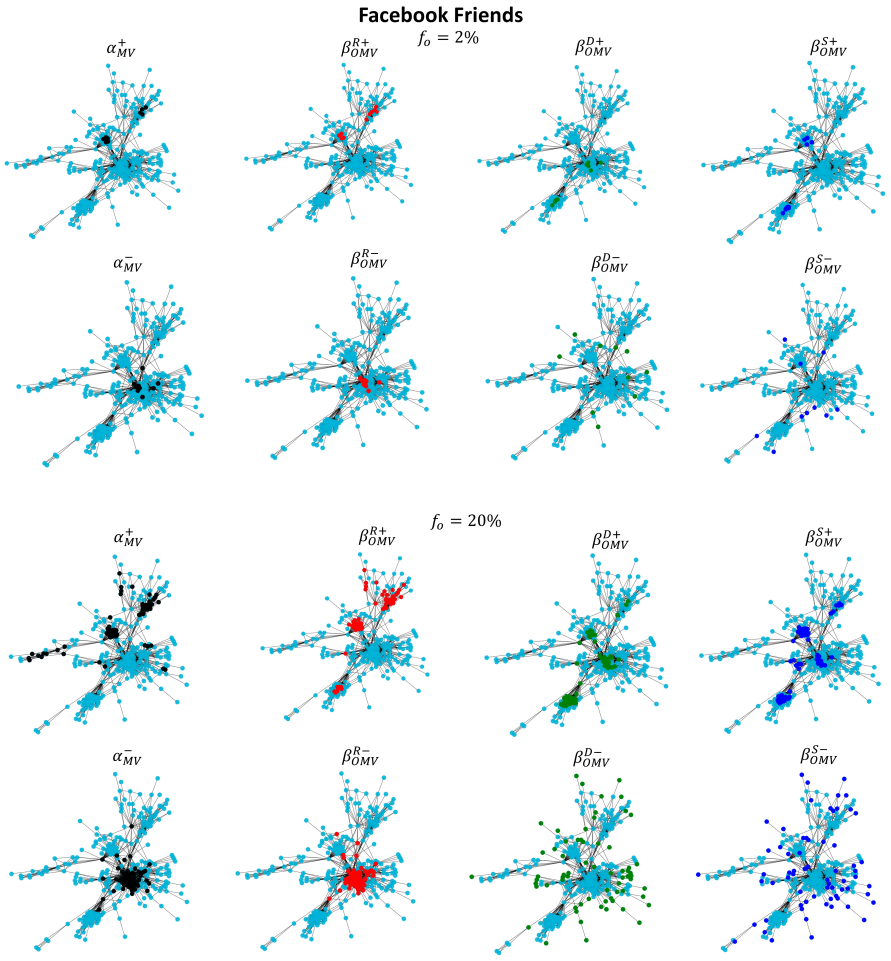


Figure 5.9: **The Facebook Friends network with the nodes chosen to be initially infected by the vitality measures**: Modularity Vitality (α_{MV}), and the three different versions of Overlapping Modularity Vitality, namely: reciprocity membership (β^{R}_{OMV}), degree membership (β^{D}_{OMV}), and node similarity (β^{S}_{OMV}). The measures use a hubs-first ranking scheme denoted by a "+" sign and a bridges-first ranking scheme marked by a "$-$" sign. The top two rows and the bottom two rows represent 2% and 20% of the fraction of initially infected nodes, respectively.

5.6/ CONCLUSION

This work proposes the Overlapping Modularity Vitality framework to identify critical nodes in networks with an overlapping community structure. One can use various definitions of community membership strengths in this framework. We investigate three alternatives: reciprocity membership, degree membership, and node similarity, obtaining three measures to target essential nodes. They present two main differences. First, they use more or less information about the overlaps of the communities. Reciprocity membership relies only on the number of overlapping communities of a node. Degree membership requires more knowledge, probing the total links of a node to each community to quantify the belonging strength of overlapping nodes to their communities. Finally, node similarity quantifies the similarity of the nodes' community membership strength vectors. The second difference concerns the ability to discriminate the nodes. Reciprocity membership is

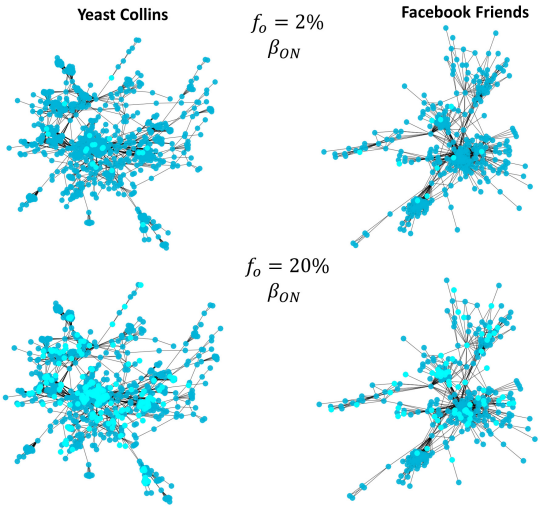


Figure 5.10: **The Yeast Collins and Facebook Friends networks with the nodes chosen to be initially infected using the OverlapNeighborhood centrality (β_{ON}).** The top and bottom rows represent 2% and 20% of the fraction of initially infected nodes, respectively.

less effective because many nodes may share the same number of communities, while node similarity and degree membership encode more nuanced differences.

One can use three ranking strategies to prioritize hubs, bridges, or both. We investigate three versions of Overlapping Modularity Vitality based on the community membership strength of the nodes (i.e., reciprocity membership, degree membership, and node similarity). We perform an extensive comparative evaluation on 21 real-world networks based on an epidemic spreading scenario using the Susceptible-Infected-Recovered epidemic model. Comparisons involve four popular overlapping centrality measures and Modularity Vitality specially designed for networks with non-overlapping community structure. Results show the effectiveness of Overlapping Modularity Vitality while marking the dependence of the performance on the resources at hand. With a limited budget of nodes (i.e., a low fraction of initially infected nodes), one should prefer Overlapping Modularity Vitality based on reciprocity membership targeting hubs or OverlapNeighborhood, a random strategy targeting the neighbors of overlapping nodes. With a higher budget of nodes, Overlapping Modularity Vitality using degree membership and node similarity with a bridges-first ranking scheme are the top measures. These results suggest that the distance between initially infected nodes is a good indicator when enough resources are available. For efficient diffusion in a multiple-spreader scenario, initially infected nodes must be far away. It is how Overlapping Modularity Vitality based on degree membership and node similarity with a bridges-first ranking scheme works. These results demonstrate the benefit of effectively integrating overlapping community structure information to identify critical nodes.

The added value of the proposed framework is its ability to integrate different structural information via a tailored overlapping modularity definition. Indeed, no universal definition suits all real-world networks' scenarios. Subsequently, an adaptable measure is required to cover multiple real-world cases. This study gives practical indications to practitioners to identify influential nodes in a network based on budget and information availability.

6

Conclusion and Future Perspectives

6.1/ Conclusion

We are surrounded by a multitude of complex networks naturally organized into communities. Depending on what is spreading within a network, diffusion dynamics can directly impact our lives, either positively or negatively. For instance, the rapid spread of information through networks can lead to the rapid dissemination of knowledge and ideas, which can drive progress and innovation. On the other hand, the rapid spread of information through networks can also have negative impacts, such as the spread of misinformation and the amplification of harmful or divisive content. It is thus crucial to gain a deeper understanding of diffusion dynamics, how community structure impacts diffusion, and which nodes play a central role in the spread of diffusion. This thesis addresses research gaps toward a better understanding of the interplay between the network's structure, the influence of important nodes and their identification, and diffusion dynamics in complex systems. It sheds light on the impact of the community structure on the dynamical spreading of diverse models, how by identifying influential nodes, we can better understand how diffusion dynamics flow through the network, and how these nodes can be harnessed to achieve desired outcomes.

To this end, in Chapter 2, we recalled the essentials for this thesis. First, we presented the centrality measures spanning from classical measures (neighborhood-based, path-based, iterative refinement-based) to community-aware measures (non-overlapping and overlapping). Then, we deliberated about diffusion models and highlighted two popular epidemic models (SI, SIR) and two information diffusion models (LT, IC). Finally, we introduced several network quality measures based on network connectivity and on models while putting more emphasis on overlapping modularity. We highlighted the flexibility of overlapping modularity in incorporating various types of information and presented two mathematical definitions in which it can be written.

In Chapter 3, driven by the reality that community structure has a confinement effect on diffusion dynamics and there are many diffusion models provided in the literature, we analyzed how the output of four diffusion models, namely the SI, SIR, IC, and LT models, depends on the network's structure and the seed nodes selected based on the community-aware centrality measures. All these models are mathematical models used to study the spread of diseases or information in a network. Each model has its own set of possible states and conditions for the spread to propagate. We showed that the strength

of the community structure and budget availability in synthetic and real-world networks significantly impact how diffusion spreads. Additionally, the SI, SIR, and IC dynamics converge while the LT dynamics diverge within a given community structure strength and budget availability. The differences between the diffusion models, mainly seen at a limited budget availability, is credited to the fact that the conditions in the SI, SIR, and IC models are well suited to select bridge-like nodes as it is easier for the epidemic or information to circulate from one community to another compared to the LT model. Studying these models and their relationship with seed nodes and the network's community structure is important because they help us understand how diseases or information can spread through a population, and thus how we can evaluate the effectiveness of different intervention strategies and make informed decisions about how to amplify the spread of positive diffusion mechanisms.

In Chapter 4, aroused by the revelation that centrality measures tend to target influential nodes contiguous to each other, we proposed a community-aware ranking scheme that naturally gravitates towards selecting nodes separated from each other by capitalizing on the network's community structure. Given any network and the rankings of any centrality measure, the proposed ranking scheme cycles through the network's communities from the largest to the and guarantees that every community has its most influential node targeted for diffusion maximization. This way, the ranking scheme avoids the diminishing return of the influential nodes' influence. It also ensures that all the network regions are not left intact. The proposed ranking scheme was compared against the traditional descending order ranking scheme using six centrality measures on synthetic and real-world networks. Results show that the proposed ranking scheme maximizes diffusion at a larger scale than the descending order ranking scheme, notably when the network has a strong community structure strength and a large number of communities of heterogeneous sizes. Moreover, the proposed ranking scheme is independent of the network type (whether directed or undirected, weighted or unweighted) and requires no additional information other than the network's communities and the rankings of any centrality measure.

In Chapter 5, we address the problem of identifying influential nodes in networks with overlapping community structure. Although they excel, most community-aware centrality measures consider non-overlapping community structures. Additionally, measures adapted for overlapping communities are hardcoded and not flexible for varying or missing information. We propose the "Overlapping Modularity Vitality" framework to resolve these issues. It identifies essential nodes based on their contribution to the network's overlapping modularity. It allows targeting top hubs or bridges or simultaneously both types of nodes. We use three alternative definitions of overlapping modularity to investigate this framework (reciprocity membership, degree membership, and node similarity). We perform extensive simulations based SIR model in an epidemic spreading process scenario. Results show that the proposed measures outperform their non-overlapping counterpart and prominent overlapping centrality measures reported in the literature. Knowing that we are always confronted with inconsistent and missing data, the proposed framework offers versatility and flexibility to any overlapping information about the node(s) and can run whether the information is fully available or not.

6.2/ FUTURE PERSPECTIVES

The analysis and propositions done in this thesis open numerous future perspectives. They can be divided into two main categories: extensions or new open questions that can be raised for further investigations. In the following, we outline these future perspectives for every contribution presented in the thesis in a non-exhaustive manner.

We first start with the investigation of the diffusion dynamics, community structure influence, and seed nodes selected based on community-aware centrality measures presented in Chapter 3. Several research works include dynamic community detection algorithms [142–144]. One interesting work encompasses investigating how the alteration in the communities within a network over time impacts the dissemination of diseases, information, or other diffusive phenomena. In other words, how the change in communities can help or hinder the diffusion of these processes. Another work could include considering both the network structure and the diffusion dynamics of the processes being studied to develop new algorithms for identifying the most influential nodes in a network. One can also explore the spread of diseases/information in networks comprising several layers, such as networks of social and professional contacts, and how these multiplex networks affect the diffusion of diseases/information. Finally, one can examine more sophisticated models that can include exogenous events, such as the introduction of a vaccine or the emergence of a new information source, and how these exogenous events affect the diffusion of diseases, information, or other processes starting at specific seed nodes in a network.

We now refer to the proposition of the community-aware ranking scheme given in Chapter 4. The proposed ranking scheme in its current form is not well-adapted for networks with few and large communities. Indeed, it is susceptible to target influential nodes in the subsequent iterations nearby each other as the ranking scheme does not have many communities to iterate over. One possible way to overcome this problem is to impose a distance condition. If the iterator goes back to the community it already visited, the next node to be selected must not be less than a distance of a specific value set depending on the diameter of the community. However, this method requires more information about the network, and thus it is more computationally demanding if the network is large. Another approach could be inverting the iterator after every complete run until the budget is reached. That is, selecting the most influential nodes in the first iteration and then selecting the least influential nodes in the second iteration, to augment the chances of selecting nodes not in the same region. One could also further divide a given community to sub-communities and then select the most influential nodes within these sub-communities. This approach allows us to zoom in to the network at the mesoscale level for a finer selection of influential nodes that are not in the same region. An alternate avenue of research would involve tailoring the ranking method to fit multilayer and temporal networks.

We now discuss the possible future perspectives related to the Overlapping Modularity Vitality framework presented in Chapter 5. One can consider extending this work in several directions. First, one approach could be to explore alternative methods for assessing the quality of the community structure in the framework instead of relying on overlapping modularity. These could include measures based on network connectivity or models. In this manner, it would be valuable to examine the extent to which each node contributes to and influences the chosen quality measure to understand better the node's overall impact on the network's overlapping community structure. Doing so gives us a new perspective

on the node's influence, which could help maximize its impact. This can be followed up by intersecting the results obtained using different quality measures, diffusion models, and machine learning to predict the node's diffusion influence. One way to quantify the node's diffusion influence under a given diffusion model is to set the in question node only as active/infected and run the diffusion model. After the model reaches the steady state, the total number of active/infected nodes acts as an indicator of the node's influence throughout the whole network. Another research work is related to improving the scalability of the framework. The framework could be optimized to handle large-scale networks. Moreover, one can explore the possible extensions of the proposed framework to multilayer and temporal networks.

BIBLIOGRAPHY

[1] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[2] Pratha Sah, Lisa O Singh, Aaron Clauset, and Shweta Bansal. Exploring community structure in biological networks with random graphs. *BMC bioinformatics*, 15(1):1–14, 2014.

[3] Jon Kleinberg and Steve Lawrence. The structure of the web. *Science*, 294(5548):1849–1850, 2001.

[4] John Hopcroft, Omar Khan, Brian Kulis, and Bart Selman. Tracking evolving communities in large linked networks. *Proceedings of the National Academy of Sciences*, 101(suppl_1):5249–5253, 2004.

[5] Mrinmaya Sachan, Danish Contractor, Tanveer A Faruquie, and L Venkata Subramaniam. Using content and interactions for discovering communities in social networks. In *Proceedings of the 21st international conference on World Wide Web*, pages 331–340, 2012.

[6] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.

[7] Linyuan Lü, Duanbing Chen, Xiao-Long Ren, Qian-Ming Zhang, Yi-Cheng Zhang, and Tao Zhou. Vital nodes identification in complex networks. *Physics Reports*, 650:1–63, 2016.

[8] Carla Sciarra, Guido Chiarotti, Francesco Laio, and Luca Ridolfi. A change of perspective in network centrality. *Scientific reports*, 8(1):1–9, 2018.

[9] Ahmed Ibnoulouafi, Mohamed El Haziti, and Hocine Cherifi. M-centrality: identifying key nodes based on global position and local degree variation. *Journal of Statistical Mechanics: Theory and Experiment*, 2018(7):073407, 2018.

[10] Zi-Ke Zhang, Chuang Liu, Xiu-Xiu Zhan, Xin Lu, Chu-Xu Zhang, and Yi-Cheng Zhang. Dynamics of information diffusion and its applications on complex networks. *Physics Reports*, 651:1–34, 2016.

[11] Marshall H Becker. Sociometric location and innovativeness: Reformulation and extension of the diffusion model. *American sociological review*, pages 267–282, 1970.

[12] Christian Pescher, Philipp Reichhart, and Martin Spann. Consumer decision-making processes in mobile viral marketing campaigns. *Journal of interactive marketing*, 28(1):43–54, 2014.

[13] Francesco Bonchi, Carlos Castillo, Aristides Gionis, and Alejandro Jaimes. Social network analysis and mining for business applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):1–37, 2011.

[14] Guilherme Ferraz De Arruda, André Luiz Barbieri, Pablo Martin Rodriguez, Francisco A Rodrigues, Yamir Moreno, and Luciano da Fontoura Costa. Role of centrality for the identification of influential spreaders in complex networks. *Physical Review E*, 90(3):032812, 2014.

[15] Aram Galstyan and Paul Cohen. Cascading dynamics in modular networks. *Physical Review E*, 75(3):036109, 2007.

[16] Leon Danon, Alex Arenas, and Albert Díaz-Guilera. Impact of community structure on information transfer. *Physical Review E*, 77(3):036103, 2008.

[17] Azadeh Nematzadeh, Emilio Ferrara, Alessandro Flammini, and Yong-Yeol Ahn. Optimal network modularity for information diffusion. *Physical review letters*, 113(8):088701, 2014.

[18] László Lovász. Random walks on graphs. *Combinatorics, Paul erdos is eighty*, 2(1-46):4, 1993.

[19] Pascal Pons and Matthieu Latapy. Computing communities in large networks using random walks. In *International symposium on computer and information sciences*, pages 284–293. Springer, 2005.

[20] Roger Guimera and Luis A Nunes Amaral. Functional cartography of complex metabolic networks. *nature*, 433(7028):895–900, 2005.

[21] Zhiying Zhao, Xiaofan Wang, Wei Zhang, and Zhiliang Zhu. A community-based approach to identifying influential spreaders. *Entropy*, 17(4):2228–2252, 2015.

[22] Naveen Gupta, Anurag Singh, and Hocine Cherifi. Centrality measures for networks with community structure. *Physica A: Statistical Mechanics and its Applications*, 452:46–59, 2016.

[23] Shi-Long Luo, Kai Gong, and Li Kang. Identifying influential spreaders of epidemics on community networks. *arXiv preprint arXiv:1601.07700*, 2016.

[24] Muluneh Mekonnen Tulu, Ronghui Hou, and Talha Younas. Identifying influential nodes based on community structure to speed up the dissemination of information in complex network. *IEEE Access*, 6:7390–7401, 2018.

[25] Zakariya Ghalmane, Mohammed El Hassouni, and Hocine Cherifi. Immunization of networks with non-overlapping community structure. *Social Network Analysis and Mining*, 9(1):45, 2019.

[26] Thomas Magelinski, Mihovil Bartulovic, and Kathleen M Carley. Measuring node contribution to community structure with modularity vitality. *IEEE Transactions on Network Science and Engineering*, 8(1):707–723, 2021.

[27] Christopher Blöcker, Juan Carlos Nieves, and Martin Rosvall. Map equation centrality: community-aware centrality based on the map equation. *Applied Network Science*, 7(1):1–24, 2022.

[28] Laurent Hébert-Dufresne, Antoine Allard, Jean-Gabriel Young, and Louis J Dubé. Global efficiency of local immunization on complex networks. *Scientific reports*, 3(1):1–8, 2013.

[29] Manish Kumar, Anurag Singh, and Hocine Cherifi. An efficient immunization strategy using overlapping nodes and its neighborhoods. In *Companion Proceedings of the The Web Conference 2018*, pages 1269–1275, 2018.

[30] Fatemeh Taghavian, Mostafa Salehi, and Mehdi Teimouri. A local immunization strategy for networks with overlapping community structure. *Physica A: Statistical Mechanics and its Applications*, 467:148–156, 2017.

[31] Zakariya Ghalmane, Chantal Cherifi, Hocine Cherifi, and Mohammed El Hassouni. Centrality in complex networks with overlapping community structure. *Scientific reports*, 9(1):1–29, 2019.

[32] Stephany Rajeh, Marinette Savonnet, Eric Leclercq, and Hocine Cherifi. Identifying influential nodes using overlapping modularity vitality. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 257–264, 2021.

[33] Mark S Granovetter. The strength of weak ties. In *Social networks*, pages 347–367. Elsevier, 1977.

[34] Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3-5):75–174, 2010.

[35] José Roberto C Piqueira, Manuel AM Cabrera, and Cristiane M Batistela. Malware propagation in clustered computer networks. *Physica A: Statistical Mechanics and its Applications*, 573:125958, 2021.

[36] Matthew Richardson and Pedro Domingos. Mining knowledge-sharing sites for viral marketing. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 61–70, 2002.

[37] Rahul Goel, Loïc Bonnetain, Rajesh Sharma, and Angelo Furno. Mobility-based sir model for complex networks: with case study of covid-19. *Social Network Analysis and Mining*, 11(1):1–18, 2021.

[38] Damon Centola and Michael Macy. Complex contagions and the weakness of long ties. *American journal of Sociology*, 113(3):702–734, 2007.

[39] Douglas Guilbeault, Joshua Becker, and Damon Centola. Complex contagions: A decade in review. *Complex spreading phenomena in social systems*, pages 3–25, 2018.

[40] Josh A Firth. Considering complexity: animal social networks and behavioural contagions. *Trends in ecology & evolution*, 35(2):100–104, 2020.

[41] Nikolaj Horsevad, David Mateo, Robert E Kooij, Alain Barrat, and Roland Bouffanais. Transition from simple to complex contagion in collective decision-making. *Nature Communications*, 13(1):1–10, 2022.

[42] Mason A Porter and James P Gleeson. Dynamical systems on networks. *Frontiers in Applied Dynamical Systems: Reviews and Tutorials*, 4, 2016.

[43] Mahdi Jalili and Matjaž Perc. Information cascades in complex networks. *Journal of Complex Networks*, 5(5):665–693, 2017.

[44] Giulio Rossetti, Letizia Milli, Salvatore Rinzivillo, Alina Sîrbu, Dino Pedreschi, and Fosca Giannotti. Ndlib: a python library to model and analyze diffusion processes over complex networks. *International Journal of Data Science and Analytics*, 5(1):61–79, 2018.

[45] Erhu Du, Eddie Chen, Ji Liu, and Chunmiao Zheng. How do social media and individual behaviors affect epidemic transmission and control? *Science of the Total Environment*, 761:144114, 2021.

[46] Adrien Guille, Hakim Hacid, Cecile Favre, and Djamel A Zighed. Information diffusion in online social networks: A survey. *ACM Sigmod Record*, 42(2):17–28, 2013.

[47] Mei Li, Xiang Wang, Kai Gao, and Shanshan Zhang. A survey on information diffusion in online social networks: Models and methods. *Information*, 8(4):118, 2017.

[48] Yang Yujie. A survey on information diffusion in online social networks. In *Proceedings of the 2020 European Symposium on Software Engineering*, pages 181–186, 2020.

[49] William Ogilvy Kermack and Anderson G McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721, 1927.

[50] Jacob Goldenberg, Barak Libai, and Eitan Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing letters*, 12(3):211–223, 2001.

[51] Mark Granovetter. Threshold models of collective behavior. *American journal of sociology*, 83(6):1420–1443, 1978.

[52] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of ACM SIGKDD*, pages 137–146, 2003.

[53] Wei Chen, Yajun Wang, and Siyu Yang. Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 199–208, 2009.

[54] Yuchen Li, Ju Fan, Yanhao Wang, and Kian-Lee Tan. Influence maximization on social graphs: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 30(10):1852–1872, 2018.

[55] Suman Banerjee, Mamata Jenamani, and Dilip Kumar Pratihar. A survey on influence maximization in a social network. *Knowledge and Information Systems*, 62(9):3417–3455, 2020.

[56] Romualdo Pastor-Satorras, Claudio Castellano, Piet Van Mieghem, and Alessandro Vespignani. Epidemic processes in complex networks. *Reviews of modern physics*, 87(3):925, 2015.

[57] Yamir Moreno, Romualdo Pastor-Satorras, and Alessandro Vespignani. Epidemic outbreaks in complex heterogeneous networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, 26(4):521–529, 2002.

[58] Mark EJ Newman. Spread of epidemic disease on networks. *Physical review E*, 66(1):016128, 2002.

[59] Alexei Vazquez, Balazs Racz, Andras Lukacs, and Albert-Laszlo Barabasi. Impact of non-poissonian activity patterns on spreading processes. *Physical review letters*, 98(15):158702, 2007.

[60] Byungjoon Min, K-I Goh, and Alexei Vazquez. Spreading dynamics following bursty human activity patterns. *Physical Review E*, 83(3):036102, 2011.

[61] Alankar Jain, Vivek Borkar, and Dinesh Garg. Fast rumor source identification via random walks. *Social Network Analysis and Mining*, 6(1):1–13, 2016.

[62] Daryl J Daley and David G Kendall. Stochastic rumours. *IMA Journal of Applied Mathematics*, 1(1):42–55, 1965.

[63] Jiajia Wang, Laijun Zhao, and Rongbing Huang. Siraru rumor spreading model in complex networks. *Physica A: Statistical Mechanics and its Applications*, 398:43–55, 2014.

[64] Yi Jing, Liu Peiyu, Tang Xiaobing, and Liu Wenfeng. Improved sir advertising spreading model and its effectiveness in social network. *Procedia Computer Science*, 129:215–218, 2018.

[65] Mark Bognanni, Douglas Hanley, Daniel Kolliner, and Kurt Mitman. Economics and epidemics: Evidence from an estimated spatial econ-sir model. Technical report, 2020.

[66] Julia Amador. The stochastic sira model for computer viruses. *Applied Mathematics and Computation*, 232:1112–1124, 2014.

[67] Ida Kristine Fjelldal, Anna Kralj, and Brent Moyle. Profanity in viral tourism marketing: A conceptual model of destination image reinforcement. *Journal of Vacation Marketing*, 28(1):52–63, 2022.

[68] Gabriel E Kreindler and H Peyton Young. Rapid innovation diffusion in social networks. *Proceedings of the National Academy of Sciences*, 111(supplement_-3):10881–10888, 2014.

[69] H Peyton Young. The evolution of social norms. *economics*, 7(1):359–387, 2015.

[70] Damon Centola. *How behavior spreads: The science of complex contagions*, volume 3. Princeton University Press Princeton, NJ, 2018.

[71] Khubaib Ahmed Qureshi, Rauf Ahmed Shams Malick, Muhammad Sabih, and Hocine Cherifi. Complex network and source inspired covid-19 fake news classification on twitter. *IEEE Access*, 9:139636–139656, 2021.

[72] Ons Abid, Salma Jamoussi, and Yassine Ben Ayed. Deterministic models for opinion formation through communication: A survey. *Online Social Networks and Media*, 6:1–17, 2018.

[73] Paulo Shakarian, Abhivav Bhatnagar, Ashkan Aleali, Elham Shaabani, Ruocheng Guo, et al. *Diffusion in social networks*. Springer, 2015.

[74] Feng Zhou, Jianxin Roger Jiao, and Baiying Lei. A linear threshold-hurdle model for product adoption prediction incorporating social network effects. *Information Sciences*, 307:95–109, 2015.

[75] Fabián Riquelme, Pablo Gonzalez-Cantergiani, Xavier Molinero, and Maria Serna. Centrality measure in social networks based on linear threshold model. *Knowledge-Based Systems*, 140:92–102, 2018.

[76] Yifan Zhang and S Thomas Ng. Robustness of urban railway networks against the cascading failures induced by the fluctuation of passenger flow. *Reliability Engineering & System Safety*, 219:108227, 2022.

[77] Wen Xu and He Chen. Scalable rumor source detection under independent cascade model in online social networks. In *2015 11th International Conference on Mobile Ad-hoc and Sensor Networks (MSN)*, pages 236–242. IEEE, 2015.

[78] Ceren Budak, Divyakant Agrawal, and Amr El Abbadi. Limiting the spread of misinformation in social networks. In *Proceedings of the 20th international conference on World wide web*, pages 665–674, 2011.

[79] Qiyao Wang, Yuehui Jin, Tan Yang, and Shiduan Cheng. An emotion-based independent cascade model for sentiment spreading. *Knowledge-Based Systems*, 116:86–93, 2017.

[80] Thi Kim Thoa Ho, Quang Vu Bui, and Marc Bui. Homophily independent cascade diffusion model based on textual information. In *International Conference on Computational Collective Intelligence*, pages 134–145. Springer, 2018.

[81] Jaewon Yang and Jure Leskovec. Defining and evaluating network communities based on ground-truth. *Knowledge and Information Systems*, 42(1):181–213, 2015.

[82] Malek Jebabli, Hocine Cherifi, Chantal Cherifi, and Atef Hamouda. Community detection algorithm evaluation with ground-truth data. *Physica a: Statistical mechanics and its applications*, 492:651–706, 2018.

[83] Tamás Nepusz, Andrea Petróczi, László Négyessy, and Fülöp Bazsó. Fuzzy communities and the concept of bridgeness in complex networks. *Physical Review E*, 77(1):016107, 2008.

[84] Jörg Reichardt and Stefan Bornholdt. Statistical mechanics of community detection. *Physical review E*, 74(1):016110, 2006.

[85] Andrea Lancichinetti, Santo Fortunato, and János Kertész. Detecting the overlapping and hierarchical community structure in complex networks. *New journal of physics*, 11(3):033015, 2009.

[86] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006.

[87] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[88] Pasquale De Meo, Emilio Ferrara, Giacomo Fiumara, and Alessandro Provetti. Generalized louvain method for community detection in large networks. In *2011 11th international conference on intelligent systems design and applications*, pages 88–93. IEEE, 2011.

[89] Peter Ronhovde and Zohar Nussinov. Multiresolution community detection for megascale networks by information-based replica correlations. *Physical Review E*, 80(1):016109, 2009.

[90] Peter Ronhovde and Zohar Nussinov. Local resolution-limit-free potts model for community detection. *Physical Review E*, 81(4):046114, 2010.

[91] Mark EJ Newman. Fast algorithm for detecting community structure in networks. *Physical review E*, 69(6):066133, 2004.

[92] Mingming Chen and Boleslaw K Szymanski. Fuzzy overlapping community quality metrics. *Social Network Analysis and Mining*, 5(1):1–14, 2015.

[93] Shihua Zhang, Rui-Sheng Wang, and Xiang-Sun Zhang. Identification of overlapping community structure in complex networks using fuzzy c-means clustering. *Physica A: Statistical Mechanics and its Applications*, 374(1):483–490, 2007.

[94] Martin Rosvall, Daniel Axelsson, and Carl T Bergstrom. The map equation. *The European Physical Journal Special Topics*, 178(1):13–23, 2009.

[95] Alcides Viamontes Esquivel and Martin Rosvall. Compression of flow can reveal overlapping-module organization in networks. *Physical Review X*, 1(2):021025, 2011.

[96] Hao Peng, Azadeh Nematzadeh, Daniel M Romero, and Emilio Ferrara. Network modularity controls the speed of information diffusion. *Physical Review E*, 102(5):052316, 2020.

[97] Andrea Lancichinetti, Santo Fortunato, and Filippo Radicchi. Benchmark graphs for testing community detection algorithms. *Physical review E*, 78(4):046110, 2008.

[98] Stefano Battiston, Michelangelo Puliga, Rahul Kaushik, Paolo Tasca, and Guido Caldarelli. Debtrank: Too central to fail? financial networks, the fed and systemic risk. *Scientific reports*, 2(1):1–6, 2012.

[99] Stephen Morris. Contagion. *The Review of Economic Studies*, 67(1):57–78, 2000.

[100] David Easley and Jon Kleinberg. *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge university press, 2010.

[101] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[102] Sergey N Dorogovtsev and Jose FF Mendes. Evolution of networks. *Advances in physics*, 51(4):1079–1187, 2002.

[103] Kwang-Il Goh, Eulsik Oh, Hawoong Jeong, Byungnam Kahng, and Doochul Kim. Classification of scale-free networks. *Proceedings of the National Academy of Sciences*, 99(20):12583–12588, 2002.

[104] Stefano Boccaletti, Vito Latora, Yamir Moreno, Martin Chavez, and D-U Hwang. Complex networks: Structure and dynamics. *Physics reports*, 424(4-5):175–308, 2006.

[105] Dimitrios Tsiotas. Detecting differences in the topology of scale-free networks grown under time-dynamic topological fitness. *Scientific reports*, 10(1):1–16, 2020.

[106] Martin Rosvall and Carl T Bergstrom. Maps of random walks on complex networks reveal community structure. *PNAS*, 105(4):1118–1123, 2008.

[107] Karen Stephenson and Marvin Zelen. Rethinking centrality: Methods and examples. *Social networks*, 11(1):1–37, 1989.

[108] Peter Csermely, Tamás Korcsmáros, Huba JM Kiss, Gábor London, and Ruth Nussinov. Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacology & therapeutics*, 138(3):333–408, 2013.

[109] Zhen Wang, Yamir Moreno, Stefano Boccaletti, and Matjaž Perc. Vaccination and epidemics in networked populations—an introduction, 2017.

[110] Andreas Klein, Henning Ahlf, and Varinder Sharma. Social activity and structural centrality in online social networks. *Telematics and Informatics*, 32(2):321–332, 2015.

[111] Sebastian Wandelt and Xiaoqian Sun. Robustness estimation of infrastructure networks: On the usage of degree centrality. In *Proceedings of the 13th International Conference on Availability, Reliability and Security*, pages 1–7, 2018.

[112] Mahdi Jalili, Ali Salehzadeh-Yazdi, Shailendra Gupta, Olaf Wolkenhauer, Marjan Yaghmaie, Osbaldo Resendis-Antonio, and Kamran Alimoghaddam. Evolution of centrality measurements for the detection of essential proteins in biological networks. *Frontiers in Physiology*, 7:375, 2016.

[113] Kai Wang, Wei Quan, Nan Cheng, Mingyuan Liu, Yu Liu, and H Anthony Chan. Betweenness centrality based software defined routing: Observation from practical internet datasets. *ACM Transactions on Internet Technology (TOIT)*, 19(4):1–19, 2019.

[114] Kousik Das, Sovan Samanta, and Madhumangal Pal. Study on centrality measures in social networks: a survey. *Social network analysis and mining*, 8(1):1–11, 2018.

[115] Xiaoqian Sun, Volker Gollnick, and Sebastian Wandelt. Robustness analysis metrics for worldwide airport network: A comprehensive study. *Chinese Journal of Aeronautics*, 30(2):500–512, 2017.

[116] Maksim Kitsak, Lazaros K Gallos, Shlomo Havlin, Fredrik Liljeros, Lev Muchnik, H Eugene Stanley, and Hernán A Makse. Identification of influential spreaders in complex networks. *Nature physics*, 6(11):888–893, 2010.

[117] Doina Bucur. Top influencers can be identified universally by combining classical centralities. *Scientific reports*, 10(1):1–14, 2020.

[118] Alex Arenas, Albert Diaz-Guilera, and Conrad J Pérez-Vicente. Synchronization reveals topological scales in complex networks. *Physical review letters*, 96(11):114102, 2006.

[119] Xue-Qi Cheng and Hua-Wei Shen. Uncovering the community structure associated with the diffusion dynamics on networks. *J. Stat. Mech. Theory Exp.*, 2010(04):P04024, 2010.

[120] Roy M Anderson and Robert M May. Population biology of infectious diseases: Part i. *Nature*, 280(5721):361–367, 1979.

[121] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

[122] Wei Liu, Matteo Pellegrini, and Xiaofan Wang. Detecting communities based on network topology. *Scientific reports*, 4(1):1–7, 2014.

[123] Emmanuel Abbe. Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531, 2017.

[124] Silvia Bartolucci, Francesco Caravelli, Fabio Caccioli, and Pierpaolo Vivo. Emerging locality of network influence. *arXiv preprint arXiv:2009.06307*, 2020.

[125] Santo Fortunato and Darko Hric. Community detection in networks: A user guide. *Physics reports*, 659:1–44, 2016.

[126] Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *nature*, 435(7043):814–818, 2005.

[127] Stephen Kelley, Mark Goldberg, Malik Magdon-Ismail, Konstantin Mertsalov, and Al Wallace. Defining and discovering communities in social networks. In *Handbook of Optimization in Complex Networks*, pages 139–168. Springer, Boston, MA, 2012.

[128] Fergal Reid, Aaron McDaid, and Neil Hurley. Partitioning breaks communities. In *Mining Social Networks and Security Informatics*, pages 79–105. Springer, Dordrecht, 2013.

[129] Ioannis Psorakis, Stephen Roberts, Mark Ebden, and Ben Sheldon. Overlapping community detection using bayesian non-negative matrix factorization. *Physical Review E*, 83(6):066114, 2011.

[130] Jierui Xie, Stephen Kelley, and Boleslaw K Szymanski. Overlapping community detection in networks: The state-of-the-art and comparative study. *Acm computing surveys (csur)*, 45(4):1–35, 2013.

[131] Jaewon Yang and Jure Leskovec. Overlapping community detection at scale: a nonnegative matrix factorization approach. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 587–596, 2013.

[132] Vinícius da Fonseca Vieira, Carolina Ribeiro Xavier, and Alexandre Gonçalves Evsukoff. A comparative study of overlapping community detection methods from the perspective of the structural properties. *Applied Network Science*, 5(1):1–42, 2020.

[133] Samrat Gupta and Pradeep Kumar. An overlapping community detection algorithm based on rough clustering of links. *Data & Knowledge Engineering*, 125:101777, 2020.

[134] Amit Kumar, Debaditya Barman, Ritam Sarkar, and Nirmalya Chowdhury. Overlapping community detection using multiobjective genetic algorithm. *IEEE Transactions on Computational Social Systems*, 7(3):802–817, 2020.

[135] Osman Doluca and Kaya Oğuz. Apal: Adjacency propagation algorithm for overlapping community detection in biological networks. *Information Sciences*, 579:574–590, 2021.

[136] Dirk Koschützki, Katharina Anna Lehmann, Leon Peeters, Stefan Richter, Dagmar Tenfelde-Podehl, and Oliver Zlotowski. Centrality indices. In *Network analysis*, pages 16–61. Springer, Berlin, Heidelberg, 2005.

[137] Steve Gregory. Fuzzy overlapping communities in networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(02):P02017, 2011.

[138] Zakariya Ghalmane, Chantal Cherifi, Hocine Cherifi, and Mohammed El Hassouni. Centrality in complex networks with overlapping community structure. *Scientific reports*, 9(1):15, 2019.

[139] Qinna Wang and Eric Fleury. Overlapping community structure and modular overlaps in complex networks. In *Mining Social Networks and Security Informatics*, pages 15–40. Springer, 2013.

[140] Timothy C Havens, James C Bezdek, Christopher Leckie, Kotagiri Ramamohanarao, and Marimuthu Palaniswami. A soft modularity function for detecting fuzzy communities in social networks. *IEEE Transactions on Fuzzy Systems*, 21(6):1170–1175, 2013.

[141] Balázs Adamcsek, Gergely Palla, Illés J Farkas, Imre Derényi, and Tamás Vicsek. Cfinder: locating cliques and overlapping modules in biological networks. *Bioinformatics*, 22(8):1021–1023, 2006.

[142] Gergely Palla, Albert-László Barabási, and Tamás Vicsek. Quantifying social group evolution. *Nature*, 446(7136):664–667, 2007.

[143] Peter J Mucha, Thomas Richardson, Kevin Macon, Mason A Porter, and Jukka-Pekka Onnela. Community structure in time-dependent, multiscale, and multiplex networks. *science*, 328(5980):876–878, 2010.

[144] Marya Bazzi, Mason A Porter, Stacy Williams, Mark McDonald, Daniel J Fenn, and Sam D Howison. Community detection in temporal multilayer networks, with an application to correlation networks. *Multiscale Modeling & Simulation*, 14(1):1–41, 2016.

[145] David Lusseau, Karsten Schneider, Oliver J Boisseau, Patti Haase, Elisabeth Slooten, and Steve M Dawson. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. *Behavioral Ecology and Sociobiology*, 54(4):396–405, 2003.

[146] Ryan A. Rossi and Nesreen K. Ahmed. The network data repository with interactive graph analytics and visualization. In *AAAI*, 2015.

[147] Aaron Clauset, Ellen Tucker, and Matthias Sainz. The colorado index of complex networks, 2016. [Online]. Available: https://icon.colorado.edu/.

[148] Vito Latora, Vincenzo Nicosia, and Giovanni Russo. *Complex networks: principles, methods and applications*. Cambridge University Press, 2017. [Online]. Available: https://www.complex-networks.net/datasets.html.

[149] Benedek Rozemberczki and Rik Sarkar. Characteristic functions on graphs: Birds of a feather, from statistical descriptors to parametric models, 2020.

[150] Tiago P. Peixoto. The netzschleuder network catalogue and repository, 2020. [Online]. Available: https://networks.skewed.de/.

[151] Jérôme Kunegis. Handbook of network analysis [konect–the koblenz network collection]. *arXiv preprint arXiv:1402.5500*, 2014. [Online]. Available: http://konect.uni-koblenz.de.

[152] Jierui Xie, Boleslaw K Szymanski, and Xiaoming Liu. Slpa: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process. In *2011 ieee 11th international conference on data mining workshops*, pages 344–349. IEEE, 2011.

[153] Günce Keziban Orman, Vincent Labatut, and Hocine Cherifi. Comparative evaluation of community detection algorithms: a topological approach. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(08):P08001, 2012.

[154] Vinh-Loc Dao, Cécile Bothorel, and Philippe Lenca. Community structure: A comparative evaluation of community detection methods. *arXiv preprint arXiv:1812.06598*, 2018.

List of Figures

List of Tables

I

APPENDICES

DATA

Throughout this thesis, synthetic and real-world networks are used. The synthetic networks are generated by the LFR model with which several topological parameters can be controlled [97]. The real-world networks pertain to various domains and are characterized by diverse topological structures. All real-world networks can be obtained online [145–151]

SYNTHETIC NETWORKS

The LFR model [97] allows generating modular networks with controlled power-law degree (γ) and community size (θ) distributions. In addition, one can also tune the community structure strength through the so-called mixing parameter (μ). Small values of μ indicate a strong community structure with few links between communities. Weak community structures correspond to high values of μ with a high fraction of connections between communities. Throughout this thesis, studies involve simulations on a set of synthetic networks with diverse values for the mixing parameter (μ), community size distribution (θ), and degree distribution (γ). Table 1 reports these parameter values.

Table 1: Synthetic network parameters generated by the LFR model.

Network Parameter	Value
Number of nodes	2500
Average degree	8
Maximum degree	27
Exponent for degree distribution (γ)	[2, 2.7, 3]
Exponent for community size distribution (θ)	[2, 2.7, 3]
Minimum community size	4
Maximum community size	250
Mixing parameter (μ)	[0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35 0.40, 0.70]

REAL-WORLD NETWORKS

A set of 50 real-world networks is collected from different fields (social, biological, ecological, infrastructure, and collaboration networks). They are provided in Table 2.

COMMUNITY DETECTION ALGORITHMS

If the community structure of the networks under test is unknown, one uses a community detection algorithm. The non-overlapping community structure is identified using Infomap [106] and Louvain [87]. The overlapping community structure is uncovered by the Speaker-Listener Label Propagation Algorithm (SLPA) [152] and the Lancichinetti Fortunato Method (LFM) [85]. Previous research shows the effectiveness of these algorithms in uncovering communities [82, 153, 154].

Domain	Network's name and number
Animal networks	Dolphins (1), Reptiles (2)
Biological networks	Budapest Connectome (3), Blumenau Drug (4), E. coli Transcription (5), Human Protein (6), Interactome Vidal (7), Kegg Metabolic (8), Malaria Genes (9), Mouse Visual Cortex (10), Yeast Collins (11), Yeast Protein (12)
Collaboration networks	DBLP (13), AstroPh (14), C.S. PhD (15), GrQc (16), NetSci (17), New Zealand Collaboration (18)
Offline social networks	Adolescent health (19), Jazz (20), Zachary Karate Club (21), Madrid Train Bombings (22)
Infrastructural networks	EU Airlines (23), EuroRoad (24), Internet Autonomous Systems (25), Internet Topology Cogentco (26), London Transport (27), U.S. Power Grid (28), U.S. Airports (29), U.S. States (30)
Actor networks	Game of Thrones (31), Les Misérables (32), Marvel Partnerships (33), Movie Galaxies (34)
Miscellaneous networks	911AllWords (35), Bible Nouns (36), Board of Directors (37), DNC Emails (38), Football (39), Polbooks (40)
Online social networks	DeezerEU (41), Ego Facebook (42), Facebook Friends (43), Facebook Organizations (44), Caltech (45), Facebook Politician Pages (46), Hamsterster (47), PGP (48), Princeton (49), Retweets Copenhagen (50)

Table 2: The fifty real-world networks used in this study divided into eight different domains.

Infomap is based on data compression and random walks. A random walker is first deployed on the network. A random walker is more likely to stay in the same community than leave for another community due to the modular structure of real-world networks. Huffman coding is used to retain the information about staying and leaving the communities using prefix codes (for the communities) and suffix codes (for the nodes). The suffix codes can be used several times. By compressing the random walk description, the non-overlapping community structure is revealed.

Louvain is based on optimizing the modularity of a network. It consists of two iterative steps. First, communities are uncovered by maximizing local modularity between the nodes. Second, a new network is built consisting of nodes as the communities found in the previous step, and modularity is maximized on this new network until no further gain can be achieved.

Speaker-Listener Label Propagation Algorithm (SLPA) is based on information diffusion. In SLPA, an extended version of the Label Propagation Algorithm, nodes can save their gained knowledge (i.e., different labels) from previous iterations. Initially, each node belongs to a different community. In other words, each node has a unique label. Next, a random node is selected as a listener. Labels are propagated from their speakers (i.e., neighbors). The process keeps iterating based on a user-defined number of iterations T. When the diffusion of labels stops, the memory of each node resembles a probability distribution of membership strengths to different communities. A probability r is then assigned to transform the membership strengths into binary memberships. In this thesis, T

is set to 20, and r is set to 0.01.

Lancichinetti Fortunato Method (LFM) is based on the fitness function. The fitness function quantifies the strength of the community structure through the internal and total degrees of communities. It incorporates a resolution parameter for detecting overlapping and hierarchical communities simultaneously. Low values of the resolution parameter yield few but large communities, while high values produce numerous small communities. One starts with any node at random considered as a community. The optimization process adds nodes to this community, maximizing the fitness function. After reaching a local maximum, one chooses another node randomly from the unassigned nodes. Then, the process iterates until all nodes belong to at least one community. Nodes assigned to communities can also be added to newly created communities, resulting in an overlapping community structure. In this thesis, the resolution parameter is set to 0.8.

RELATIVE OUTBREAK SIZE

The outbreak size, which indicates the total number of nodes in the recovered or activated state, is calculated for each centrality measure when the diffusion process ends. This value indicates the spreading/influence ability of the centrality measure at a given fraction of initially infected/active nodes f_o. The higher this value, the more effective the centrality measure in selecting the most influential nodes. The relative difference in the outbreak size is defined as follows:

$$\Delta R = \frac{R_c - R_r}{R_r} \tag{1}$$

where:

- R_c denotes the outbreak size using a centrality measure c under test
- R_r denotes the outbreak size using the reference centrality

ΔR is positive if the centrality measure under test is more effective than the reference. Else, ΔR is negative. It is important to note that in the SIR model, ΔR has the same meaning as ΔA in the IC model, which is the relative size of activation. These measures both assess the number of nodes that become infected or activated at the end of the diffusion process.

Table 3: The Degree centrality of each node in the toy network with their respective descending order and community-aware ranks.

Node ID	Degree	Descending order ranks	Community-aware ranks
1	5	2	4
2	3	12	16
3	3	13	18
4	5	3	7
5	7	1	1
6	4	4	10
7	4	5	13
8	2	17	20
9	2	18	21
10	2	19	22
11	2	20	15
12	3	14	12
13	4	6	3
14	4	7	6
15	4	8	9
16	2	21	17
17	3	15	11
18	4	9	5
19	4	10	2
20	3	16	14
21	4	11	8
22	2	22	19

Table 4: The Betweenness centrality of each node in the toy network with their respective descending order and community-aware ranks.

Node ID	Betweenness	Descending order ranks	Community-aware ranks
1	0.039	14	18
2	0.056	13	16
3	0.200	5	1
4	0.099	11	13
5	0.158	7	4
6	0.100	10	10
7	0.146	8	7
8	0.010	19	21
9	0.014	17	20
10	0.000	20	22
11	0.016	16	15
12	0.038	15	12
13	0.287	1	3
14	0.266	3	9
15	0.280	2	6
16	0.000	21	17
17	0.012	18	14
18	0.080	12	11
19	0.125	9	8
20	0.190	6	5
21	0.221	4	2
22	0.000	22	19

Table 5: The number of communities and their minimum and maximum sizes for the real-world networks based on communities identified by Infomap.

Network	Number of communities	Minimum Size	Maximum Size
EU Airlines	10	2	332
Ego Facebook	72	2	471
U.S. Airports	39	2	226
Facebook Friends	21	2	72
Facebook Politician Pages	180	2	421
Madrid Train Bombings	5	3	38
Yeast Collins	61	2	119
Malaria Genes	11	2	86
NetSci	38	3	32
Reptiles	55	2	42
Marvel Partnerships	27	2	11
911AllWords	842	2	7609
U.S. Power Grid	422	3	44
Board of Directors	78	5	23
PGP	896	2	160
Princeton	29	2	3714
London Transport	50	4	14
EuroRoad	111	3	22
Internet Topology Cogentco	27	4	19
DNC Emails	38	2	231
Yeast Protein	164	2	49
Blumenau Drug	6	3	29
Retweets Copenhagen	92	3	32
Hamsterster	64	2	692
Human Protein	99	2	645
Caltech	11	2	270
Facebook Organizations	51	11	526
Interactome Vidal	222	2	124
AstroPh	675	2	547
DeezerEU	1395	2	446
DBLP	376	2	528
Adolescent Health	136	3	237
Bible Nouns	88	3	131

Table 6: The number of communities and their minimum and maximum sizes for the real-world networks based on communities identified by Louvain.

Network	Number of communities	Minimum Size	Maximum Size
EU Airlines	8	27	103
Ego Facebook	15	19	548
U.S. Airports	12	2	137
Facebook Friends	10	5	92
Facebook Politician Pages	29	17	585
Madrid Train Bombings	5	3	22
Yeast Collins	22	3	130
Malaria Genes	8	6	71
NetSci	18	6	56
Reptiles	19	6	55
Marvel Partnerships	14	6	19
911AllWords	14	5	2052
U.S. Power Grid	41	26	241
Board of Directors	26	13	69
PGP	101	6	672
Princeton	11	4	1752
London Transport	17	11	31
EuroRoad	23	20	84
Internet Topology Cogentco	12	6	27
DNC Emails	10	2	210
Yeast Protein	33	9	84
Blumenau Drug	5	12	19
Retweets Copenhagen	23	7	85
Hamsterster	13	6	307
Human Protein	14	32	604
Caltech	9	9	164
Facebook Organizations	11	35	1267
Interactome Vidal	34	4	412
AstroPh	32	5	1629
DeezerEU	91	4	4326
DBLP	22	5	1933
Adolescent Health	19	16	358
Bible Nouns	17	6	247

Document generated with LaTeX and:
the LaTeX style for PhD Thesis created by S. Galland — http://www.multiagent.fr/ThesisStyle
the tex-upmethodology package suite — http://www.arakhne.org/tex-upmethodology/