Non-assortative Community Structures in Complex Networks

XUANCHI LIU

Supervisor: Eduardo G. Altamann
Associate Supervisor: Tristram J. Alexander

A thesis submitted in fulfilment of
the requirements for the degree of
Doctor of Philosophy

School of Mathematics
Faculty of Science
The University of Sydney
Australia

24 October 2024

Declaration of Authorship

I, Xuanchi Liu, declare that this thesis titled, “Non-assortative Community Structure in Complex Networks”, and the whole work presented in it are my own. I want to confirm that:

- This work was done wholly while in candidature for a research degree at the University of Sydney
- Where any part of this thesis has not previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main source of help.
- Where the thesis is based on work done myself jointly with others, I have made clear exactly what was done by others and what I have contributed my-self.

Sincerely,

Xuanchi Liu

29/06/2024

Authorship Attribution Statement

Chapter 3 of this thesis is published as

Cathy Xuanchi Liu, Tristram J Alexander, and Eduardo G Altmann. Nonassortative relationships between groups of nodes are typical in complex networks. PNAS nexus, 2(11):pgad364, 2023.

Chapter 4 of this thesis has been submitted for publication. A pre-print is available at

Cathy Xuanchi Liu, Tristram J Alexander, and Eduardo G Altmann. A generative model for community types in directed networks. preprint arXiv:2405.14168.

For these two publications, I, Xuanchi Liu, developed the methodologies, performed the computations and statistical analysis, and interpreted the analysis, and wrote the drafts of the manuscripts.

In addition to the statements above, in cases where I am not the corresponding author of a published item, permission to include the published material has been granted by the corresponding author.

Xuanchi Liu
Signature:
Date:

As supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Eduardo G Altmann	Tristram J Alexander
Signature:	Signature:
Date:	Date:

Abstract

The identification of mesoscale structures is essential to understand the organization of complex networks. Despite the recognition of various structural patterns in many networks, previous works on the detection of groups of users have focused on finding specific types, such as assortative and core-periphery structures, and developing computational methods to find these mesoscale structures within a network. In this thesis, we go beyond the problem of detecting a specific type of mesoscale arrangement and instead ask what type of mesoscale structures exist, how typical each is, and how we can interpret them. We start with introducing a methodology that is able to identify and systematically classify all possible community structure types in directed multi-graphs, based on the pairwise relationship between groups. The data survey on 55 empirical networks shows that other structural patterns are typical in networks. A particularly prevalent new type of relationship, which we call a source-basin structure, reveals a new type of information flow from a sparsely connected group of nodes (source) to a densely connected group (basin). To gain a better understanding of the formation of these structures, we introduce and investigate a generative model of network evolution with two communities that reproduces all four pairwise community types that exist in directed networks: assortative, core-periphery, disassortative, and the newly introduced source-basin type. With node connectivity changing during evolution, we find that a difference in assortative preference of the two communities is a necessary condition to obtain a core-periphery relationship and that a difference in the average in-degree of the communities is a necessary condition for a source-basin relationship. We further apply the methods developed to social media data and examine the significant role that community structures play in networks. Notably, core-periphery or source-basin structures are obtained in almost every network, confirming the widespread appearance of non-assortative structures in social networks. A detailed study of source-basin structures reveals a new pattern of information flow, where more central nodes having few interactions with each other influence

an actively interacting group. It also shows the potential of hitherto unidentified relationships to explain the organization of complex networks.

Acknowledgements

I would like to express my deepest gratitude to all those who have given me enormous support throughout my PhD candidature, without whom the completion would not be possible.

First and foremost, I want to thank my supervisor, Eduardo G. Altmann, who guided me with inexhaustible patience and supported me in every way. His encouragement and recognition of my work gave me the courage to finish this project. I also want to thank my co-supervisor, Tristram J. Alexander, who introduced me to the fascinating field of data science, shared the Twitter dataset, and dedicated his time to reviewing my work. I extend my sincere appreciation to Monika Bednarek and the authors of Ref. [Bed+22], who shared social media data and brought meaningful interpretations from social and linguistic perspectives; Mike Thelwall, who shared SentiStrength implementations; Eckehard Olbrich and the authors of Ref [Gai+21], who shared Twitter datasets about Germany political debates. Additionally, I am grateful to Samuel Isaac Jelbart, Georg Gottwald, and Mikhail Prokopenko for offering me the opportunity to participate in the SDG workshop and C3-2023, where I gained new ideas and different perspectives for this thesis.

I am deeply grateful to my colleagues and friends in Eduardo's group—Karina Arias Calluari, Oscar Fajardo, and other members — for their friendship, collaboration, and intellectual exchange, which filled my academic journey with joy and enriched my research with fruitful discussions and shared experiences.

I am immensely thankful to my friends, Emily Wang, who cheered me up whenever I was upset and invited me to enjoy the beautiful views of Australia; Yuechen Wang, with whom I have been friends since high school and had engaging discussions on common interests; and Bochen Wu, my best friend from childhood who gave me unconditional support. I also offer my thanks to my friends Heshang, Aza, and Xingxing, who sincerely supported and encouraged me even though we had never met in person.

I would like to express my profound gratitude to my family for their unwavering love, encouragement, and understanding throughout this challenging yet rewarding journey. Their unconditional support and belief in my abilities have been a constant source of motivation and inspiration. Special thanks to my adorable puppy, Oreo, who is the greatest gift of my life and has brought me immense happiness.

Last but not least, I would like to express my sincere gratitude to the Faculty of Science Research Stipend Scholarship for their generous financial support, which has been instrumental in the successful completion of this PhD thesis.

Contents

CHAPTER 1

Introduction

In the digital era, social media like Facebook, X(former Twitter), and Instagram, have become an increasingly popular way that enables users to express their thoughts, receive information, as well as to communicate, and collaborate with each other. With a vast amount of information generated by users, social media provides many opportunities for the study of human interaction, information dissemination, and community formation. These platforms are characterized by not only a tremendous amount of connections but also multiple types of interactions and content. For example, users can retweet, reply, or mention other users with texts, photos, or even videos. These characteristics make social media a valuable information source that offers rich context for studying complex systems and understanding human behavior at scale.

Networks have become a popular approach to analyze social media data where each individual can be represented by a node, and the interaction between individuals can be represented by an edge [New18; Bar13]. This approach is used in the new field, known as computational social science [Laz+09; Con+12], takes advantage of the scalability of the network approach and goes beyond the traditional social analysis, which is limited to smaller datasets. Extensive research has been conducted in this area such as political polarization [Cin+21; CM15], information diffusion [Gle+14], and opinion mining [Mit+13; Cod+15].

The analysis of networks has traditionally been conducted at the micro level, focusing on individual nodes and their local properties such as degree and neighbour, and at the macro level, examining the network as a whole and considering global properties such as clustering coefficient [New18; Bar13]. However, the microscale analysis often becomes computationally intensive for networks with more than a few nodes, and the macroscale analysis tends to

lose detailed information about the network's structure. To balance the detail and overview, mesoscale analysis based on groups of nodes provides an intermediate level of detail and offers insights into the fundamental properties of underlying social organization and online interaction. Previous analysis of mesoscale structures considered how different communities of nodes relate to each other and has led to the development of various types of analysis, including assortative structure where nodes are strongly connected within communities [For10; OM+09]; core-periphery, in which information flows from a tightly connected core community of users to a loosely connected periphery [BE00; ZMN15; Yan+18a]; hierarchical structure where smaller groups are contained within larger ones [CMN08], and bow-tie structure with a strongly connected core, an in-periphery influencing core and a out-periphery receiving influence from core [YZN11] etc. Classic examples of well-known assortative structures include the famous Zachary's karate club [Zac77], scientific collaboration [Red98], and ecosystem [BU89]. Core-periphery structures were also obtained in many fields such as social [Yan+18b] and biological [BBB18] networks.

The structure of communities plays a vital important role in understanding the underlying organization of the network and gaining insights into group-level relationships in networks. Community structure is often related to the problem of finding communities, noted as community detection in network science [New18; GSA07; For10; FN22; Liu+20]. A substantial amount of research has been conducted on this problem, providing mathematically and computationally achievable algorithms. In particular, most approaches aim for assortative communities with some representative algorithms such as modularity maximisation [CNM04; Blo+08], spectral clustering [Von07] and infomap [RB08]. Some algorithms are designed to identify core-periphery structures [BE00; Rom+17; Ell+20; KM17]. Apart from methods that are targeted for specific types of mesoscale structure, an approach based on statistical inference of network structure, which is represented through probabilistic models such as the Stochastic Block Model (SBM) has gained increasing attention recently, leaving community identification to the data to determine the most significant statistical signature that leads to the clustering of nodes into blocks (communities) [HLL83; KN11; Pei19].

Despite significant advancements in community detection algorithms, the interpretation and explanation of these communities remain challenging and underdeveloped. While various methods can effectively partition networks into subgroups, understanding the underlying structure of these communities is still elusive. The exploration of community structure is not merely an academic pursuit but holds significant implications for various real-world applications such as political voting [Muc+10], social relationships [Red+11] and biological studies [Lew+10; BBB18]. This gap between the ability to detect communities and the ability to explain their formation and significance underscores the need for new theoretical and computational frameworks that can explain the structural patterns hidden in the network. It motivates us to explore the field of community structure to understand the organizational patterns that drive the functionality and behavior of various systems. The problem we address in this thesis is to quantify and interpret the existence of different mesoscale structures between pairwise communities which we call as community structures. We develop mathematical models and computational tools to obtain a deeper understanding of the presence and origin of different community structures in networks.

We start with a review of graph theory, community structure analysis, existing community detection methodologies, and network evolution models in Chapter 2. In Chapter 3, we focus on the identification and classification of possible types of community structure between pairwise communities using undirected and directed multi-graphs. In particular, we find a new type, source-basin, where information flows from a sparsely connected source community of users to a strongly connected basin. In real-world network data, we find the dominance of assortative structure but also the ubiquitous appearance of non-assortative structures. To gain a deeper understanding of underlying mechanisms of the formation of these structures, we propose a network evolution model that shows the the emergence of community structure types. In Chapter 4, we focus on directed networks and propose a simple generative model that changes the connectivity of nodes to observe all structure types. We find the parameters and conditions needed for the appearance of each community type. In Chapter 5, we apply the methods developed in previous chapters to social media data and examine the significant role that community structures play in the organization of networks. We take retweet and reply user interactions from X focusing on three political debate events in Australia and Germany.

Notably, core-periphery or source-basin structures are identified in almost all networks. In particular, we find that source-basin structures revreal a new type of information flow winthin networks that is different from classical core-periphery and play an important role in the organization of user interaction on online social platforms.

CHAPTER 2

Review of existing methods

In this chapter, we review previous results that will in our analysis including graph theory, community structures, and methods that are targeted for specific structures. Then, we explain several representative, traditional and new community detection methods, including Louvain, Spectral Clustering, Infomap, Stochastic Block Model and Deep Neural Network approach, that have been used in this project. We also review social media data, and community detection works based on it. Lastly, we review the concept of the network evolution model and introduce several fundamental models.

2.1 Graph Theory

In recent years, graph theory has established itself as an important mathematical tool in a wide variety of subjects, ranging from operational research and chemistry to genetics and linguistics and from electrical engineering and geography to sociology and architecture [Wes+01]. In network science [New18; Bar13], the complex relationships between objects can be represented as abstract networks with nodes and links. A node corresponds to an object such as a human in social networks, a protein in protein-protein networks, and a city in flight networks. The link between nodes corresponds to relationships such as the friendship between two people, protein interaction, or a flight route. We introduce graph notations used throughout our work.

Let $G = (V, E)$ be a graph with node set $V = \{v_1, ..., v_n\}$ and e_{ij} be a link (or edge) from node v_i to v_j which can be directed ($e_{ji} \neq e_{ij}$) or undirected ($e_{ji} = e_{ij}$). Multiple edges between the same nodes may occur, which can be represented as multiplicity $w_{ij} \in \mathbb{N}$ of

e_{ij}. The edge that starts and ends at the same node is called the self-edge. A graph with no multi-edges and no self-edge is called a simple graph, and a graph with multi-edges is called a multi-graph.

The fundamental mathematical representation of a network is the adjacency matrix defined as $n \times n$ matrix $A = \{a_{ij}\}$, $i, j = 1, ..., n$ with

$$a_{ij} = \begin{cases} 1 & \text{, if an edge exists from node } v_i \text{ to } v_j \\ 0 & \text{, otherwise} \end{cases} \tag{2.1}$$

in simple graphs, and

$$a_{ij} = \begin{cases} w_{ij} & \text{, if an edge exists from node } v_i \text{ to } v_j \\ 0 & \text{, otherwise} \end{cases} \tag{2.2}$$

in multi-graphs.

The degree of a node v_i is the number of edges connected to it which is defined as

$$k_i = \sum_{j=1}^{n} a_{ij} \tag{2.3}$$

For directed graphs, each node v_i has three degree quantities: 1). in-degree which counts the number of edges with node v_i as end and is calculated as $k_i^{\text{in}} = \sum_{j=1}^{n} a_{ji}$; 2). out-degree which counts the number of edges starting from node v_i and is calculated as $k_i^{\text{out}} = \sum_{j=1}^{n} a_{ij}$; 3). total degree which is the sum of in and out degree $k_i^{\text{total}} = k_i^{\text{in}} + k_i^{\text{out}}$. The degree matrix D is defined as the diagonal matrix with the degrees $k_1, ..., k_n$ on the diagonal.

Another matrix representation of graph is the Laplacian matrix which is defined as

$$L = D - A \tag{2.4}$$

with D and A from Eqs. (2.1),(2.2),(2.3). The Laplacian matrix is a fundamental tool in graph theory and network analysis, and is used to study the structural properties of graphs. It plays

a crucial role in various applications, including spectral graph theory, random walks, and network dynamics [New18].

While most studies of networks employed an abstraction in which a single type of relationship is considered in systems, the complexity of real-world data often requires a more realistic framework for the systems that objects interact with each other in complicated patterns that can encompass multiple types of relationships [Kiv+14]. For instance, flights between two cities are from multiple airlines [Car+13], and two humans can be friends and colleagues [RD03]. The " multilayer " feature is introduced to study complex systems such as sociology and computer science. The multilayer network can be defined as a sequence of graphs $\{G_\alpha\} = \{V_\alpha, E_\alpha\}$, $\alpha = \{1, ..., b\}$. Usually, the node sets are the same across the different layers (i.e., $V_\alpha = V_\beta$ for all α, β), and multiple types of edges exist for different layers. In particular, this approach can be beneficial for the study of social media data, which we will describe in Sec. 2.3.

2.2 Community Structure

We are mostly interested in the analysis of network conducted on the intermediate level where groups of nodes (or communities) share the same connectivity pattern. It benefits us by providing insights into how nodes group together and interact, which is often overlooked in micro or macro analysis. Classic examples of structures include assortative structure [For10; OM+09], core-periphery [BE00; ZMN15; Yan+18a], hierarchical structure [CMN08], and bow-tie structure [YZN11]. The emergence of these structures has shown in networks such as assortative communities in karate club [Zac77], core-periphery in neuro network [BBB18] and other non-assortative structures in various networks [MPA24; Moo+11; Pee15; Pee17] as a result of interactions between nodes, and plays an important role in affecting the behaviour of complex systems [BE00; Van+14]. As networks continue to grow in size and complexity, mesoscale structure is an essential tool for understanding network organization and unlocking the secrets of interconnected systems across various domains. In this project, we focus on

relationships between pairwise communities, which we call community structure. Here, we review two widely used structures in social analysis: assortative and core-periphery.

Assortative

Assortative community structure in network science refers to the organization of nodes within a network into communities where nodes within the same community are more densely connected to each other than to nodes in other communities [For10; OM+09]. The property is shown in many networks, and the concept is now widely used in complex networks of social, biological, technological, or informational fields, representing the nature of modular connectivity patterns [AG05; Con+11a; Pow+11; GN02]. Many algorithms and methods have been developed to detect assortative structures, ranging from classical techniques like spectral clustering [Von07] and modularity optimization [CNM04; Blo+08] to more advanced approaches incorporating dynamic processes taking place on networks [RB08] that we will explain in the following section.

Core-Periphery

Core-periphery is a common notion in social analysis that refers to a specific organizational pattern within a network where nodes are divided into two distinct groups: the core and the periphery [BE00]. The core-periphery structure represents a network pattern that has a densely connected core and a sparsely connected periphery influenced by the core. Another popular hierarchy structure of core-periphery is identified by k-core decomposition that finds the largest subset of nodes in k-shell such that every node has at least k connections to other nodes in that subset [GYW21]. The periphery is in the outer shells and the cores are in the inner shells. In this project, we will focus on the two-block model of core-periphery and not discuss the k-core decomposition. The concept of core-periphery structure has now been widely applied in many fields, such as social, biological and transportation, and reveals important structure properties [Yan+18b; BBB18; Ver+16]. The intuitive ideas of core-periphery used in 1990s research or before include a cohesive subgroup that cannot be further divided or two groups of nodes the core group is densely connected, and the periphery is sparsely connected. A definition of core-periphery was given by Borgatti and Everett in 1999. They proposed

discrete and continuous models for core-periphery [BE00]. Based on this model, many approaches have been proposed to detect core-periphery structure [Rom+17; Ell+20; KM17].

2.3 Community Detection

The study of community structure has a long history and is closely related to the ideas of community detection in graph theory [GN02]. Community detection is a problem of allocating users into groups according to their connectivity in such a way that nodes with similar connectivity belong to the same group. We review five representative community detection methods from different perspectives.

2.3.1 Modularity Maximisation

From social network analysis, a good group partition divides nodes into densely connected subgroups, that is, assortative structure. To measure partition quality, Newman and Girvan proposed a quality function called modularity [New06]. It is defined as the difference between the fraction of edges within groups and the expected fraction in a randomized version of the network.

We consider two nodes i and j with degree k_i and k_j respectively from a random undirected network. Each edge starting from node i has equal probability links to $2m-1$ rest ending point. Since node j has k_j stubs, the probability of one edge from node i link to node j is $\frac{k_j}{2m-1}$. The expected number of links between i and j is $\frac{k_i k_j}{2m-1}$. For large network, we use the approximate expression $2m$ in the denominator. The difference is then calculated as

$$a_{ij} - \frac{k_i k_j}{2m}$$

Summing over all node pairs gives the equation for modularity

$$Q = \frac{1}{2m} \sum_{i,j} (a_{ij} - \frac{k_i k_j}{2m}) \delta(g_i, g_j)$$

where g_i, g_j is the community assignment of node i, j.

The calculation of modularity gives a scalar value between -1 and 1, with a higher value meaning a "better" community partition. By maximizing this objective function, we try to find groups of nodes that have maximum edges within groups, that is, assortative communities as described in Sec. 2.2. Many efforts have been made to search for assortative structure, and a large number of algorithms have been proposed, most notably the Louvain method [CNM04; Blo+08]. Introduced by Blondel et al. in 2008, the Louvain algorithm offers fast and accurate estimates of a network's assortative structure. The algorithm assumes we start with a weighted network G that has a partition that each node belongs to a different community. The algorithm is as follows:

(1) In the first phase, we consider node i and its neighbour j and move node i from its community to the community of its neighbour if the moving can increase modularity. Otherwise, we keep node i in its community. We do this process repeatedly and sequentially for all nodes until the modularity cannot increase by this process. Then the first phase is complete, and we start the second phase.
(2) In the second phase, we build a community network in which the nodes are the communities of network G, and the weights of links are the sum of the weight of the links between nodes in the corresponding two communities. Once we have the community network, we can repeat our first phase in this new network. The loop stops when modularity can not be improved.

The Louvain algorithm is highly regarded for its speed and scalability, making it suitable for large networks with millions of nodes and edges [LF09; YAT16]. Its ability to uncover assortative community structures and its practical efficiency has made it one of the most widely used methods in complex networks and various application domains, from social network analysis to biological network research [Cin+21; FWW17].

2.3.2 Spectral Clustering

Spectral clustering algorithms aim to recover the hidden structured groups in the observed networks that have a minimum number of edges between groups and a maximum number of edges within groups [Von07]. Spectral clustering is based on local neighbour relationships and detects communities by the properties of eigenvalues and eigenvectors of normalized graph Laplacian matrix L, which is defined in Eq. (2.4).

The Laplacian matrix has a nice property that the smallest eigenvalue of L is 0, and the corresponding eigenvector is the constant one vector $\mathbb{1}$. The unnormalized graph Laplacian and its eigenvalues and eigenvectors can be used to describe many properties of graphs [Bol91]. One important example that can be used in spectral clustering is the connection between the eigenvalue and the number of connected components.

Proposition from Ref. [Von07]: Let G be an undirected graph with non-negative weights. Then the multiplicity k of the eigenvalue 0 of L equals the number of connected components $C_1, ..., C_k$ in the graph. The eigenspace of eigenvalue 0 is spanned by the indicator vectors $\mathbb{1}_{C_1}, ..., \mathbb{1}_{C_k}$ of those components.

The clustering problem in graph G with partition $\{g_1 = \{v_{1_1}, ...\}, ..., g_k = \{v_{k_1}, ...\}\}$ turns out to be minimization of cut-edges (edge between communities), which is defined as

$$\min \text{cut}(g_1, ..., g_k) := \frac{1}{2} \sum_{i=1}^{k} A(g_i, \bar{g_i})$$

where $A(g_i, g_j) := \sum_{v_i \in g_i, v_j \in g_j} a_{ij}$ is the sum of the possibly weighted edges between group g_i and g_j and $\bar{g_i}$ is the complement of subset g_i.

However, the solution of $\min$cut leads to the separation of individual nodes, which is not satisfactory [Von07]. To overcome the problem of group size, Ratiocut is proposed, taking account group size into objective function as

$$\min \text{Ratiocut}(g_1, ..., g_k) := \frac{1}{2} \sum_{i=1}^{k} \frac{A(g_i, \bar{g}_i)}{|g_i|}$$

where $|g_i|$ is the number of nodes in group g_i. The objective function takes a small value if the size of the group $|g_i|$ is not large and balances the linking behaviour and group size. The problem can be rewritten as

$$\min_f f^T L f$$
$$\text{subject to } f \perp \mathbf{1} \, \|f\| = \sqrt{n}$$
$$\text{where } f_{i,j} := \begin{cases} \frac{1}{\sqrt{|g_j|}} & \text{if } v_i \in g_j \\ 0 & \text{otherwise} \end{cases}$$

This is a discrete optimization problem as the entries of the solution vector f are only allowed to take discrete values, and the problem is NP-hard [Von07]. We use the relaxed versions of this problem that f_i takes arbitrary values in $\mathbb{R}$ [DCT19]. The relaxed problem becomes:

$$\min_{F \in \mathbb{R}^{n \times k}} \text{Tr}(F' L F)$$
$$\text{subject to } F'F = I$$

Rayleigh-Ritz theorem tells us that the solution is given by choosing F as the matrix, which contains the first k eigenvectors of L as columns and uses the k-means algorithm on the rows of the eigenvectors into k groups g'. Then, the communities of nodes are classified as

$$\begin{cases} v_i \in g & \text{if } f_i \in g' \\ v_i \in \bar{g} & \text{if } f_i \notin g' \end{cases}$$

The standard spectral algorithms proved to be efficient in speed and analytical tractability for dense graphs where the average node degree scales with the size of the network [DCT19]. But real networks typically are sparse that the average degree k is a constant and does not scale with the size of the network. To overcome the problem of sparsity, instead of the Laplacian

matrix, one important matrix is introduced: the $n \times n$ Bethe-Hessian H_r with $r \in \mathbb{R}$ [DCT19; SKZ14]. The Bethe-Hessian matrix with parameter $r = \zeta$ can detect communities in sparse and heterogeneous graphs. The Bethe-Hessian matrix is defined as

$$H_r = (r^2 - 1)I_n + D - rA, r \in \mathbb{R}$$

where $|r| > 1$ is a regularizer that is set to a well-defined value $|r| = r_c$ depending on the graph.

Proposition from Ref. [DCT19]: In the sparse regime, there exists a value $\zeta \leq \sqrt{c\Phi}$ for which the components of the "second" eigenvector of H_ζ align to the community labels. The clustering process is similar to the standard one: we compute the eigenvectors associated with the negative eigenvalues of both H_r and H_{-r}, and cluster them with a standard clustering algorithm such as k-means. Another advantage is that the number of communities can be informed by the number of negative eigenvalues so that no prior information of k is needed.

2.3.3 Infomap

Instead of finding groups from the composition of network represented by adjacency matrix, another motivation is to find how the information flows within the network such that information tends to stay within communities rather than between communities. If we treat network as a map of random walks which simulate the flow of information or movement through the network, communities correspond to regions where random walkers are more likely to stay for longer periods. Focusing solely on interconnection between nodes can give an incomplete picture of the network dynamics, such as how frequently a node is visited, the role a node plays in the information flow, or how certain nodes act as bridges between communities [FN22; Ros+19].

Developed by Martin Rosvall and Carl T. Bergstrom in 2008, the Infomap algorithm became an influential method for detecting communities in complex networks [RB08]. This approach uses Markovian diffusion to model information flows of random walks in the network and find system description of the probability flow which can be identified as groups of nodes.

The flow of information is represented with directed weighted edge of network. In order to capture dynamic behaviour of the system from information flow, the model use random walks as a proxy since it only requires and uses information representation in the network.

To encode the path from random walks, Huffman coding [Sma+08] is used to give names to nodes where similar nodes are given a short codeword and rare nodes are given a long codeword. The length of codeword is then linked with the description of random walk using Shannon's source coding theorem [Sha48]. The average number of bits needed to describe a single step in the random walk is bounded from below by the entropy $H(P)$ with P being the distribution of visit frequencies to the nodes. This lower bound on code length is defined as L. To capture important structure pattern of system, the model uses two levels of description, one for clusters of the network and one for individual nodes within clusters. The partitioning problem is then transferred as finding effective two descriptions defined as

$$\min L(M) = qH(Q) + \sum_{i=1}^{M} p^i H(P)$$

where $H(Q)$ is the entropy of cluster names and $H(P)$ is the entropy of within cluster movements. q and p^i are the probabilities of switching between clusters and moving within clusters respectively.

The minimization problem is, however, infeasible even in small networks. Thus, greedy search algorithms [CNM04; WT07] are used to find possible partitions. The result is refined by a simulated annealing approach [GN05] with the heat-bath algorithm.

2.3.4 Stochastic Block Model

Apart from finding cohesive groups, grouping nodes with similar behaviour is another intuitive motivation for the partitioning problem. Nodes are similar if they have similar connection patterns with other nodes. The stochastic Block model(SBM) is proposed to detect communities within networks by partitioning nodes into blocks or communities based on their connection patterns [HLL83; KN11; Pei19]. In SBM, the probability of an edge existing

between any two nodes depends on the communities to which these nodes belong, allowing for distinct intra-community and inter-community connection probabilities. This model provides a flexible framework for uncovering hidden community structures and has potential of finding other structural patterns other than assortative and core-periphery in complex networks. The associated inference techniques can be used to detect community structures in networks. In this project, we use the inference approach proposed by Peixoto [Pei19; Pei21; Pei14a; Pei17], which has available code implementation in package [Pei20] and whose core ideas are reviewed below.

SBM assumes N nodes in the network are divided into B blocks $b_i \in \{1, ..., B\}$. The linking probabilities between two nodes i and j only depend on their block memberships. Nodes in one block are statistically equivalent, meaning that they have the same linking probabilities as group edge probability, which is defined as

$$p_{ij} = p_{b_i, b_j}$$

where p_{ij} is the probability that an edge present between node i and j and p_{b_i,b_j} is the probability that an edge present between block b_i and b_j. With edge probability and block assignment available, networks can be constructed by sampling edges between nodes with conditional probability $P(\mathbf{A}|\mathbf{p}, \mathbf{b})$.

The inference procedure is reversing the generative process in that we have an adjacency matrix a_{ij} and aim for partition $\mathbf{b}$ that generates this $\mathbf{A}$. This can be formulated as finding $P(\mathbf{b}|\mathbf{A})$. From Bayes theorem, the probability can be written as

$$P(\mathbf{b}|\mathbf{A}) = \frac{P(\mathbf{A}|\mathbf{b})P(\mathbf{b})}{P(\mathbf{A})} \tag{2.5}$$

where the marginal likelihood integrated over the remaining model parameters $P(\mathbf{A}|\mathbf{b})$ is

$$P(\mathbf{A}|\mathbf{b}) = \int P(\mathbf{A}|\lambda, \mathbf{b})P(\lambda|\mathbf{b})d\lambda$$

where $\lambda = \{\lambda_{rs}\}$ is average number of edges existing between any two nodes belonging to group r and s. The evidence of the network is

$$P(\mathbf{A}) = \sum_{\mathbf{b}} P(\mathbf{A}|\mathbf{b})P(\mathbf{b})$$

which is the total probability over all possible $\mathbf{b}$. $P(\mathbf{b})$ and $P(\lambda|\mathbf{b})$ in the above equations are the prior distributions that represent our knowledge about the model before we observe any data. Many algorithms have been proposed to solve the inference problem with different approximation methods and priors.

The choice of prior distributions prompted to be "uninformative" when we have no prior knowledge about the dataset so that each possible parameter combination should have an equal probability of being chosen. With uninformative priors of block and edge probability, the likelihood of the network can be re-interpreted as the joint likelihood of a microcanonical model given by

$$P(\mathbf{A}|\mathbf{b}) = P(\mathbf{A}|\mathbf{e}, \mathbf{b})P(\mathbf{e}|\mathbf{b})$$

where $\mathbf{e} = \{e_{rs}\}$ is the matrix of edge counts between groups. Inferring partitions from posterior distribution of Eq. 2.5 using a non-parametric approach allows us finding not only statistical significant partitions but also the optimal number of groups B which is an outcome of the inference procedure.

The above posterior probability Eq. (2.3.4) can be reframed to descriptions of network by information theory. Using the optimal lossless coding schemes like Huffman's algorithm, the asymptotic amount of information necessary to describe a variable is $-\log_2 P(x)$(in bits) [Mac03]. Then, the numerator of the posterior distribution in Eq. (2.3.4) can be written as

$$P(\mathbf{A}|\mathbf{b})P(\mathbf{b}) = P(\mathbf{A}|\mathbf{e}, \mathbf{b})P(\mathbf{e}|\mathbf{b})P(\mathbf{b}) = 2^{-\Sigma}$$

where

$$\Sigma = -\log_2 P(\mathbf{A}|\mathbf{e}, \mathbf{b}) - \log_2 P(\mathbf{e}, \mathbf{b})$$

is called the description length of the data. It measures the asymptotic amount of information necessary to encode the data $\mathbf{A}$ together with the model parameters $\mathbf{e}$ and $\mathbf{b}$. Therefore, the model that maximizes the posterior distribution will minimize the description length.

With this interpretation, we can see how the Bayesian approach prevents overfitting. As the number of groups increases, the information needed to describe the model $-\log_2 P(\mathbf{A}|\mathbf{e},\mathbf{b})$ will decrease, and the information to transmit the parameters $-\log_2 P(\mathbf{e},\mathbf{b})$ grows since more parameters are used. Thus, the parameter information term can prevent overly complex models.

Building on the basis of rigorous mathematical formulation, SBM has the benefit of providing a statistical assessment of partition results. We can also gain information on the block membership of each node by computing the probability of a node belonging to a block [Pei19]. With uninformative priors, the structural pattern that can be found by the model is not restricted to a specific structure type, while methods introduced above (modularity maximization, spectral clustering, and infomap) only look for specific structures. It is remarkable that we can be convinced that any hidden structure discovered by SBM is from data, not from detection methods.

Degree-corrected Stochastic Block Model

The shortcoming of standard SBM is that nodes in one block are indistinguishable. In particular, nodes will be grouped to have similar degrees. But it is an unrealistic behaviour, and many real networks have very heterogeneous degrees [New18]. To address this problem, Karrer and Newman proposed a modified model, degree-corrected SBM [KN11]. In the degree-corrected model, an extra set of parameters θ is added to each node to represent its expected degree, and the edge probability p_{ij} now depends on the block membership of nodes and their degrees

$$p_{ij} \sim \theta_i \theta_j p_{b_i,b_j}$$

The marginal likelihood of network in Peixoto's model takes the degree parameters k into account and is calculated as the joint likelihood of a microcanonical model given by

$$P(\mathbf{A}|\mathbf{b}) = P(\mathbf{A}|\mathbf{k},\mathbf{e},\mathbf{b})P(\mathbf{k}|\mathbf{e},\mathbf{b})P(\mathbf{e}|\mathbf{b})$$

and the description length is calculated as

$$\Sigma = -\log_2 P(\mathbf{A}|\mathbf{k},\mathbf{e},\mathbf{b}) - \log_2 P(\mathbf{k}|\mathbf{e},\mathbf{b}) - \log_2 P(\mathbf{e}|\mathbf{b})$$

Hierarchical Stochastic Block Model (hSBM)

Though the standard SBM model has advantages in many aspects, it has a resolution limit of $B_{\max} \propto \sqrt{N}$, which means the maximum number of groups can be found by SBM is bounded by B_{max} even data is generated with block number larger than this boundary value [Pei17; Pei13]. This problem is solved by including a hierarchical layer [Pei14b]. In the model proposed by Peixoto, after fitting SBM to adjacency a_{ij}, the partition result $\mathbf{b}$ can be used to construct a block network where nodes are blocks and multi-edges are the edges between blocks. The edge count between block nodes is the edge multiplicity between blocks. The block network $\mathbf{e}$ can be treated as a multi-graph and described by another SBM. The partition result of $\mathbf{e}$ can be used to construct another block network, and another SBM can be fitted. Doing this repeatedly, a hierarchical SBM with hierarchical levels of partition can be obtained.

The number of groups, the partition and the matrix of edge counts in level $l = \{0, ..., L\}$ are denoted as B_l , $\mathbf{b}_l$ and $\mathbf{e}_l$. The posterior of hSBM is defined as

$$P(\{\mathbf{b}_l\}|\mathbf{A}) = \frac{P(\mathbf{A},\{\mathbf{b}_l\})}{P(\mathbf{A})} = \frac{P(\mathbf{A},\{\mathbf{b}_l\},\{\mathbf{e}_l\})}{P(\mathbf{A})}$$

and the description length can be presented as

$$\sum = -\log_2 P(\mathbf{A}|\{\mathbf{b}_l\},\{\mathbf{e}_l\}) - \log_2 P(\{\mathbf{b}_l\},\{\mathbf{e}_l\})$$

Now, the maximum number of groups can be found by hSBM scales as $B_{\max} \propto \frac{N}{\log N}$, which is much larger than the previous limit. Another benefit of nested SBM is that it provides different scales of description of data which is helpful for real network analysis.

2.3.5 Deep Neural Networks for learning graph representation (DNGR)

With the fast and wide development of deep learning, solving community detection in deep learning techniques received increasing attention, taking advantage of dealing with massive and complex data, which is also very beneficial in network analysis. Many deep learning-based methods have been proposed in recent years [Liu+20; Su+22]. Based on the different deep neural networks used, these methods can be divided into convolutional networks, graph attention networks, generative adversarial networks, and autoencoders. In this project, we reviewed previous methods and found that most methods require additional attributes, which are trained to find structural patterns in the model [Su+22; Roz21]. However, traditional community detection problem usually finds communities based only on connectivity information and nothing else. The additional information requirement makes most deep learning-based methods not fit our problem. Moreover, a lack of open-source coding increases the difficulty of applying deep learning methods. Here, we take a survey of 63 state-of-art methods from Refs. [Su+22; Roz21] and list all methods in Table. 2.1. We find that eight of them can fit to traditional community detection problem that does not require additional metadata and performs unsupervised learning. Among these eight methods, only four have open source coding. Due to the easy implementation of Deep Neural Networks for learning graph representation(DNGR) method, we use this method in the project.

Method	**Additional Feature**	**Learning**	**Open Source**
Xin *et al.*	Not required	Surpervised	✗
SparseConv	Not required	Surpervised	✗
SparseConv2D	Not required	Semi-surpervised	✗
ComNet-R	Not required	Surpervised	✗
LGNN	Required	Supervised	✓
MRFasGCN	Required	Semi-supervised	✗
SGCN	Required	Unsupervised	✗
NOCD	Required	Unsupervised	✓

GCLN	Required	Unsupervised	✗
IPGDN	Required	Unsupervised	✗
AGC	Required	Unsupervised	✓
AGE	Required	Unsupervised	✓
CayleyNet	Required	Semi-surpervised	✓
SENet	Required	Unsupervised	✗
CommDGI	Required	Unsupervised	✗
Zhao *et al*	Required	Unsupervised	✗
DMGI	Required	Unsupervised	✓
HDMI	Required	Unsupervised	✓
MAGNN	Required	Unsupervised	✓
HeCo	Required	Unsupervised	✓
CP-GNN	Required	Unsupervised	✓
SEAL	Required	Semi-surpervised	✓
DR-GCN	Required	Semi-surpervised	✗
JANE	Required	Unsupervised	✗
ProGAN	Required	Unsupervised	✗
CommunityGAN	Not Required	Unsupervised	✓
CANE	Not Required	Unsupervised	✗
ACNE	Not Required	Unsupervised	✗
semi-DNR	Required	Semi-supervised	✓
DNE-SBP	Required	Semi-supervised	✓
UWMNE	Required	Unsupervised	✗
sE-Autoencoder	Required	Semi-supervised	✗
DANE	Required	Unsupervised	✓
Transfer-CDDTA	Required	Unsupervised	✗
DIME	Required	Unsupervised	✓
GraphEncoder	Required	Unsupervised	✓
WCD	Required	Unsupervised	✗

DFuzzy	Not Required	Unsupervised	✗
CDMEC	Required	Unsupervised	✗
DNGR	Not Required	Unsupervised	✓
DNC	Not Required	Unsupervised	✗
GRACE	Required	Unsupervised	✗
MGAE	Required	Unsupervised	✓
GUCD	Required	Unsupervised	✗
SDCN	Required	Unsupervised	✓
O2MAC	Required	Unsupervised	✓
DAEGC	Required	Unsupervised	✗
GEC-CSD	Required	Unsupervised	✗
MAGCN	Required	Unsupervised	✗
SGCMC	Required	Unsupervised	✓
DMGC	Not Required	Unsupervised	✓
TGA/TVGA	Required	Unsupervised	✗
VGECLE	Not Required	Semi-supervised	✗
DGLFRM	Required	Semi-supervised	✓
LGVG	Required	Semi-supervised	✗
VGAECD	Required	Unsupervised	✗
VGAECD-OPT	Required	Unsupervised	✗
ARGA/ARVGA	Required	Unsupervised	✓
CDCGS	Not Required	Unsupervised	✓
DMoN	Required	Unsupervised	✓
CD-ATTACK	Required	Unsupervised	✓
GALA	Required	Unsupervised	✓
ELM-CLR	Required	Unsupervised	✓

TABLE 2.1. Deep Learning Method Table: 63 deep learning methods in community detection problem. Columns include method name (Method), if an additional feature required (Addtional Feature), clustering learning type (Learning) and open source availability (Open Source).

DNGR is proposed to capture the structural information from the network through a process in which the graph is first converted into a large collection of linear structures, and then low dimensional feature representations are learned from linear structures [CLX16]. The process includes capturing graph structural information by a random surfing model, which computes the co-occurrence probability between nodes. This co-occurrence matrix represents nodes connection behaviour, and a closer distance between nodes would have a higher co-occurrence probability. The graph representation can be represented by the positive pointwise mutual information(PPMI) matrix [LG14], which is rooted in the PMI matrix proposed by Church and Hanks [CH90]. PMI matrix measures the association between two events and computes as the probability of two events occurring together dividing by this probability would be if the events were independent.

$$\text{PMI}_{i,j} = \log(\frac{\#(i,j) \cdot |D|}{\#i \cdot \#j})$$

where $|D| = \sum_i \sum_j \#(i,j)$ and

$$\text{PPMI}_{i,j} = \max(\text{PMI}_{i,j}, 0)$$

The resulting PPMI matrix can be used as graph representation fed into a deep neural network, which is powerful in learning high dimensional representations. This method uses a stacked denoising autoencoder, which consists of two steps: 1). encoding step where input features is reduced by a function $f_{\theta_1}(\dot{})$. 2). decoding step where a reconstruction function $g_{\theta_2}(\dot{})$ is used to reconstruct feature space from latent representation gained in the encoding step. The community detection problem turns out to be finding the minimum loss function, which is defined as

$$\min_{\theta_1,\theta_2} \sum_{i=1}^{n} L(x^{(i)}, G_{\theta_2}(f_{\theta_1}(\tilde{x}^{(i)})))$$

where $\tilde{x}^{(i)}$ is corrupted input data of $x(i)$ which is PPMI matrix of the underlying network, and L is the standard squared loss. In the original paper of this method, the community detection is done by applying the clustering algorithm k-means to the resulting node representation matrix, and we keep this choice in our project [CLX16].

The advantage of this method is apparent that it does not require additional node attributes of the network. In contrast, other methods build their method based on node attributes as feature representations. Secondly, this method is unsupervised, which is well suited to a clustering algorithm. Thirdly, this method extends previous work and could deal with weighted and directed networks. The only disappointing requirement of this method is the number of communities needed to give in the K-Means procedure. In our work, we address this by applying the same choice used in another paper [RB08].

2.4 Social Media Data

Social media platforms have emerged as a global communication tool characterized by user-generated content in multiple forms, including texts, images, videos, and links; multiple types of interaction between users such as like, reply, retweet and mention; and real-time information that enables the rapid dissemination and consumption of information. In particular, the rich user interaction information from social media provides opportunities for identifying groups of users with similar interactions within large-scale social networks such as retweet multi-graphs with users as nodes and retweeting relationships as edges. This can fit into the traditional community detection problem as described in Sec. 2.3, and many algorithms have been used to uncover meaningful communities that reflect real-world social structures, such as friend circles, political groups, or colleague networks [TL11; Cin+21].

The structures of communities also help in understanding the underlying social dynamics, information diffusion, and influence patterns within the network. For example, political

polarization can be found in political retweets that a highly segregated partisan structure, with extremely limited connectivity between left- and right-leaning users [Con+11a]. Retweet interactions between users are used to find political alignment [Con+11b], explain the information dissemination between pro- and anti-vaccination users [MWW20] and detect extremist groups in Twitter [BJC17]. Previous successful findings demonstrate that community detection on social media data offers valuable insights into the underlying social structures of users.

2.5 Dynamical Network Models

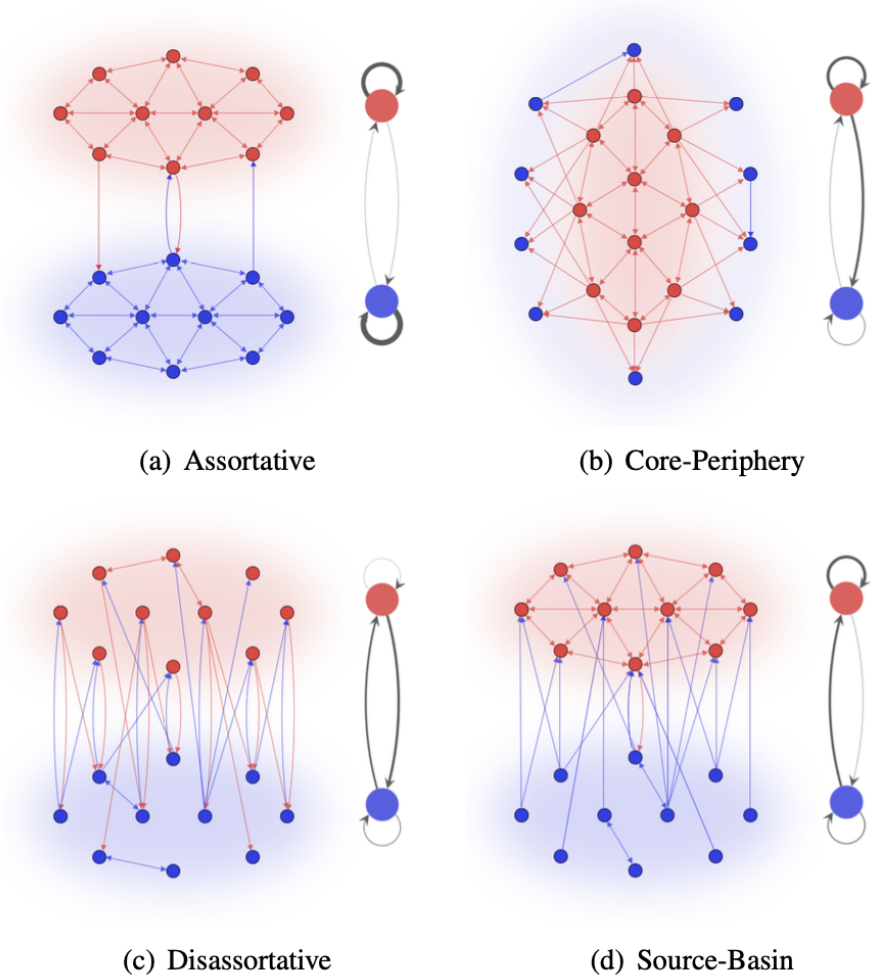
To gain a better understanding of how different network community structures emerge, we look into network models that describe the structure and dynamics of a network over time. A dynamical network model is a framework that describes how the structure or topology of networks evolve over time, allowing for the study of the mechanisms that bring particular network topologies [Say15]. Different from temporal network where the edges are time-dependent, dynamical network model focuses on the evolution of the nodes and their interactions such as the addition and removal of nodes and edges, creating a changing topology. The growth and evolution of networks have gained attention in the study of network models, explaining phenomena in various domains, such as epidemiology, neuroscience, social sciences, and ecology, where the interactions between entities and their temporal evolution play a critical role. Some important works in this field include the Small-World Model by Duncan Watts and Steven Strogatz [WS98] that captures the phenomenon where most nodes are not neighbors but can be reached from every other node by a small number of steps; Barabási–Albert Model that explains the presence of hubs or highly connected nodes in networks via mechanisms of growth and preferential attachment [AB02]; and adaptive models in which the state of nodes is related with the topology transformation of network [GDB06; GS09]. Many network evolution models have been developed for the simulation and prediction of the behaviour of real-world networks, revealing crucial insights into how complex systems function and evolve. In particular, some works show the emergence of different types of community structures, providing opportunities to gain insights about the mechanism of the formation of community structures [XS04; Ure+23; KE02].

CHAPTER 3

Classification of Community Structure

3.1 Motivation

While the problem of partitioning a network in communities has received significant attention and many methods have been developed, most algorithms are designed to identify specific structures, such as assortative and core-periphery structures [CNM04; Blo+08; Von07; RB08]. This focus limits the discovery of mesoscale structures in networks and raises questions about the relevance of specific structures within networks. For example, the extensively studied political blog network [AG05] is used to evaluate algorithms for both assortative and core-periphery structures, with different algorithms identifying each structure [KM17; Pei19]. However, compared to the development of community detection algorithms, there has been little effort to interpret mesoscale structures in the underlying network. In this chapter, we concentrate on the relationships between pairwise communities and propose an appropriate classification method for different types of community structures. We introduce a systematic classification of community structures in undirected and directed multi-graphs. Figure 3.1 illustrates the four types of structure present at the highest level of our classification in the directed case.



(a) Assortative (b) Core-Periphery

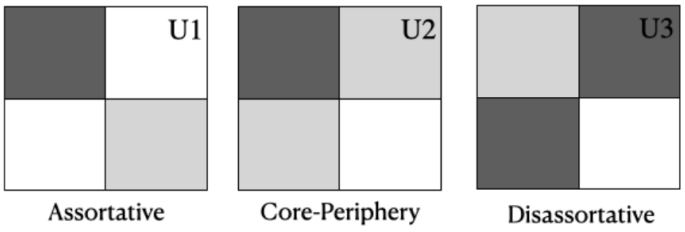
(c) Disassortative (d) Source-Basin

FIGURE 3.1. The four main types of community relationship in directed graphs: Assortative (a), Core-periphery (b), Disassortative (c), and Source-Basin (d). The left side corresponds to the graph with nodes coloured according to their community assignment and directed edges coloured according to their source nodes' colour. The right side is a graphical representation of the edge density matrix ω for the two communities.

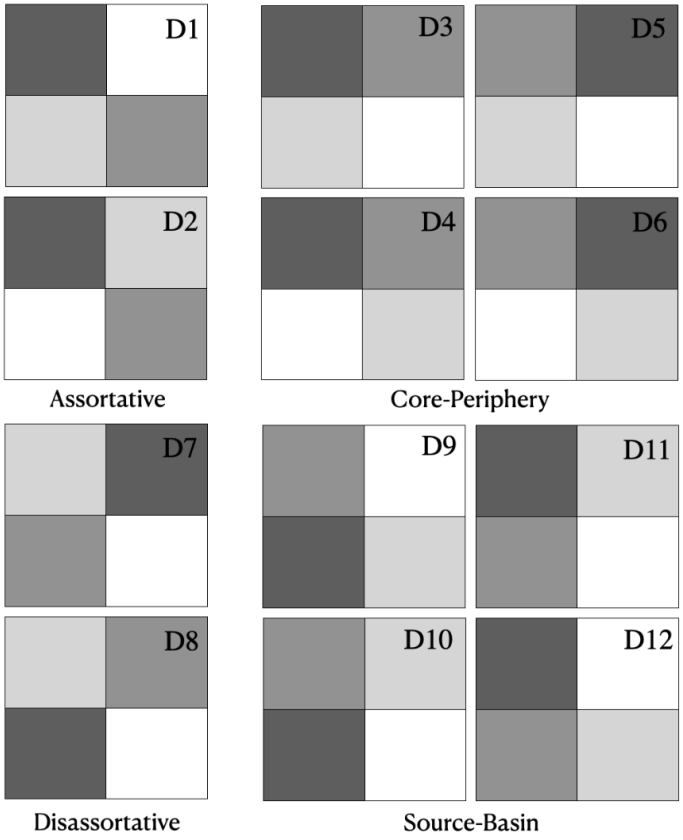
3.2 Methodology

3.2.1 Density Matrix

The community detection problem in networks corresponds to partitioning a graph G (possibly directed and weighted) with adjacency matrix $A_{ij} \in \mathbb{N}$ into $r = 1, 2, \ldots, B$ disjoint groups with N_r nodes. For a given community partition of A_{ij}, we characterise the connection



(a) Classification in undirected graph



(b) Classification in directed graph

FIGURE 3.2. Community type classification. (a) The 3 possible community types in graphs with undirected edges. (b) The 12 different possible configurations in directed graphs classified in four types. We use the convention that the link direction corresponds to the flow of information (or influence) so that a link from node i to j indicates that information is passing from i to j (or that i influences j). Assortative (Disassortative) cases have the two largest entries in the diagonal ω_{rr}, ω_{ss} (anti-diagonal ω_{rs}, ω_{sr}). The remaining 8 cases have a densely connected community (top), in which nodes are well connected to each other, and a loosely connected community (bottom). In four Core-Periphery cases the influence flows from the well-connected core to the weakly connected periphery while in the remaining four cases it flows in the reverse direction. We name this reversed core-periphery structure a source-basin structure.

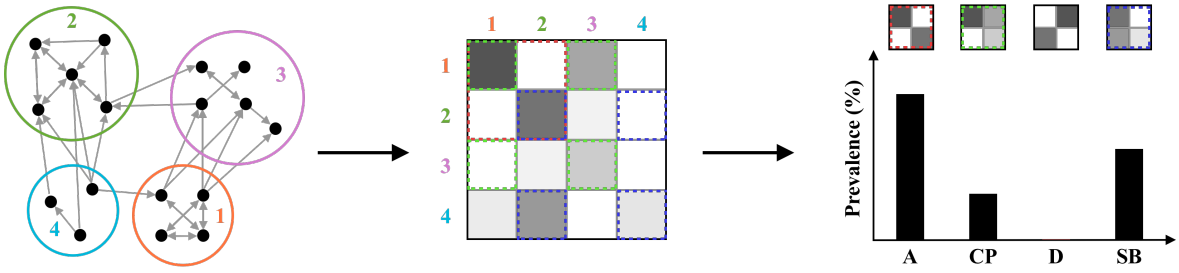


FIGURE 3.3. Illustration of our methodology for classifying community types. (a) Applying a given community detection method to a network G, we obtain a partition of the nodes in B communities ($B = 4$ in the example). (b) From the node partition we construct the $B \times B$ density matrix ω from Eqs. (3.1) or (3.2). (c) We apply the pairwise classification defined in Fig. 3.2 – and Eqs. (3.3)-(3.4) – to ω considering all pairwise combinations of communities r, s and compute the fraction of each of the four different types of community relationships, where 'A' stands for 'Assortative', 'CP' stands for 'Core-Periphery', 'D' stands for 'Disassortative' and 'SB' stands for 'Source-Basin'.

between a pair of communities r and s based on their density ω_{rs} which is given by the ratio of existing links and possible links:

$$\omega_{rs} = \begin{cases} \sum_{i\in r, j\in s} \frac{A_{ij}}{N_r N_s}, & \text{if } r \neq s \\ \sum_{i\in r, j\in s} \frac{A_{ij}}{\frac{1}{2}N_r(N_s-1)}, & \text{if } r = s \end{cases} \tag{3.1}$$

for undirected graphs ($A_{ij} = A_{ji}$) and

$$\omega_{rs} = \begin{cases} \sum_{i\in r, j\in s} \frac{A_{ij}}{N_r N_s}, & \text{if } r \neq s \\ \sum_{i\in r, j\in s} \frac{A_{ij}}{N_r(N_s-1)}, & \text{if } r = s \end{cases} \tag{3.2}$$

for directed graphs. Without loss of generality, we order the label of communities so that $\omega_{rr} \geq \omega_{ss}$ for $r < s$. Each community detection method applied to a network G will typically provide a different $B \times B$ density matrix ω and our methodology developed below applies to any method.

3.2.2 Classification

We now classify the relationship between a pair of communities r and s based on ω_{rs}. In the simplest case of undirected graphs g, there are three relevant quantities ω_{rr}, ω_{ss}, and $\omega_{rs} = \omega_{sr}$. Ranking these values by size, there are three possible configurations $U1, U2$, and $U3$ [ZMN15; BBB18], as illustrated in Fig. 3.2(a). They correspond to the two previously studied types of community structures – assortative and core-periphery – and a new case – disassortative – that has strong connections between communities (the limiting scenario $\omega_{rr} = \omega_{ss} = 0$ corresponds to a bi-partite network). A categorical classification of the pairwise interaction between communities r and s based on the ranking order of the ω_{rs} values can be written as

$$\begin{cases} \text{Assortative}, & \text{if } \omega_{rs} < \min(\omega_{rr}, \omega_{ss}), \\ \text{Core-periphery}, & \text{if } \min(\omega_{rr}, \omega_{ss}) < \omega_{rs} < \max(\omega_{rr}, \omega_{ss}), \\ \text{Disassortative}, & \text{if } \omega_{rs} > \max(\omega_{rr}, \omega_{ss}). \end{cases} \tag{3.3}$$

The case for directed graphs G leads to a wider variety of possible configurations between communities r and s. Following the same approach employed for undirected graphs, we obtain 12 possible configurations $D1, D2, \ldots, D12$ based on the ranking of the four quantities $\omega_{rr}, \omega_{rs}, \omega_{sr}$, and ω_{ss}. Figure 3.2(b) shows these twelve cases, grouped into 4 types. Configurations with the same general interpretation are grouped into the same type, which can be one of the three types observed in undirected graphs (Assortative, Core-Periphery, and Disassortative) or one new type. In particular, in the new type of relationship – denoted Source-Basin – influence flows from loosely connected source nodes to a basin of interconnected nodes. Contrary to the core-periphery relationship, the most influential and central nodes (in the source community) do not form a core as they are more connected to nodes in the other community (the basin) than to nodes in their own community.

The categorical classification of the pairwise interaction between communities r and s into different community structure types based on ω for a directed graph can be written as

$$\begin{cases} \text{Assortative}, & \text{if } \max(\omega_{rs}, \omega_{sr}) < \min(\omega_{rr}, \omega_{ss}), \\ \text{Core-Periphery}, & \text{if } \min(\omega_{rr}, \omega_{rs}) > \max(\omega_{sr}, \omega_{ss}), \\ \text{Disassortative}, & \text{if } \min(\omega_{rs}, \omega_{sr}) > \max(\omega_{rr}, \omega_{ss}), \\ \text{Source-Basin}, & \text{if } \min(\omega_{rr}, \omega_{sr}) > \max(\omega_{rs}, \omega_{ss}). \end{cases} \tag{3.4}$$

The Assortative (Disassortative) interaction is thus defined naturally as having more links to other nodes in the same (other) communities. Core-Periphery and Source-Basin are defined as interactions that show a mixed behaviour, with one sparser community more linked to the other community and one denser community more linked to itself. The difference between the two types of interactions is that in the Core-Periphery the links (influence/information) across communities flow preferentially from the denser to the sparser community, while in the Source-Basin interaction the flow is in the reversed direction.

3.2.3 Prevalence

The final step of our methodology is to apply the pairwise classification we just introduced to networks containing $B > 2$ communities. Our approach is to look at each of the possible pairwise interactions, classify its structure type $\tau \in$ {"Assortative","Core-Periphery","Disassortative","Source-Basin"} and measure the prevalence of different community structure types, as described in Fig. 3.3.

Consider a $B \times B$ edge density matrix ω obtained applying a community-detection method to a network g. The $M_{B\times B}$ interaction classification matrix of ω contains the community structure type τ of all pairs of communities $r, s \in [0, B]$, with $\tau \in$ {'Assortative', 'Core-Periphery', 'Disassortative', 'Source-Basin'}, obtained using the classification introduced in Fig. 3.2 and Eqs. (3.3)-(3.4). The fraction of τ of a network g is then given by counting how often τ appears in M as

$$P_\tau = \frac{1}{B(B-1)} \sum_{r \neq s} \delta_{\tau, M_{r,s}}, \tag{3.5}$$

where $\delta_{x,y} = 1$ if $x = y$ and $\delta_{x,y} = 0$ if $x \neq 0$.

Knowing P_τ for all structure types τ in a network g, we define the dominant type τ_d as the τ with largest P_τ and occurring types τ_o as the set of τ's with $P_\tau > 0$. We then compute the *dominance* of each type τ over a set of networks g as the fraction of networks that has dominant type $\tau_d = \tau$ as

$$\text{Dominance}(\tau) = \frac{1}{|\mathbf{g}|} \sum_{g} \delta_{\tau,\tau_d}, \tag{3.6}$$

where $|\mathbf{g}|$ is the number of networks in g. Similarly, the *occurrence* of type τ is computed as the fraction of networks in which τ occurs ($\tau \in \tau_o$) as

$$\text{Occurrence}(\tau) = \frac{1}{|\mathbf{g}|} \sum_{g} \sum_{\tau' \in \tau_o} \delta_{\tau,\tau'}. \tag{3.7}$$

In the results reported in this paper we applied the analysis above only to communities with more than 5 nodes, $N_r > 5$, because small communities play a minor role in explaining the properties of the network and the classification of their structure types τ is often not robust (when applying the tests reported in Appendix Sec. A2).

3.3 Survey

We test the methodology proposed above applying five community-detection methods to 52 real-world networks. The networks were retrieved from the Netzschleuder repository [Pei] and were selected to include both directed (26) and undirected (26) networks from a variety of domains and sizes (from 75 to 14,360 vertices and from 181 to 150,985 edges, see Appendix Sec. A1 for details). The five community-detection methods – Louvain (modularity maximization) [Blo+08; Ayn20], Infomap [RB08; Dan20], Spectral method [DCT19], degree-corrected Stochastic Block Models (SBM) [Pei19; Pei20], and DNGR (Deep Neural networks for Graph Representations) [CLX16] – were selected as representative of different approaches to this problem (see Sec. 2.3 for details and our repository [Liu23] for the numerical implementation). As pointed out in the introduction, we are particularly interested in inference based

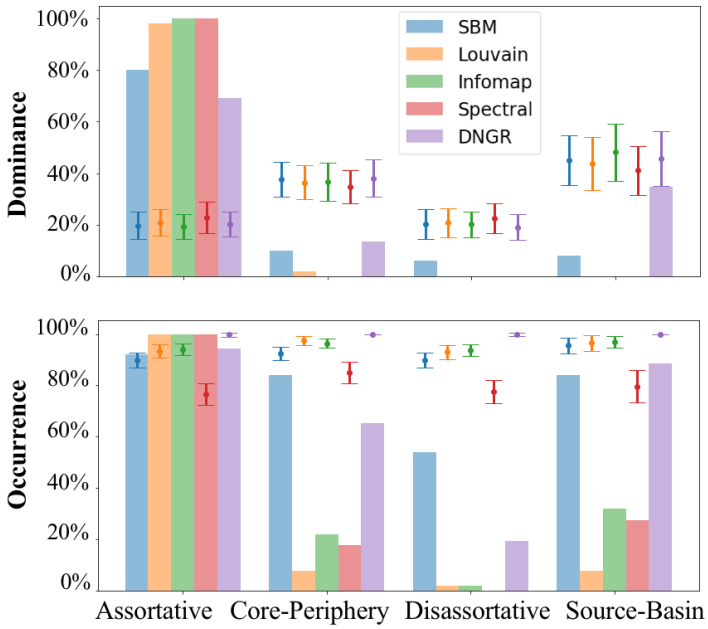


FIGURE 3.4. Community types in empirical directed and undirected networks. Top: fraction of networks (y-axis) in which each community type (x-axis) is dominant. Bottom: fraction of networks (y-axis) in which each community type (x-axis) occurs. For the Core-Periphery classification, both undirected and directed networks are considered, and for the Source-Basin classification, only directed networks are considered. See Sec. 2.3 for details on the community-detection methods and the classification is calculated by Eqs. (3.6), (3.7). The symbols with error bars were obtained in a null-model which considers a random edge-density matrix ω (see Appendix Sec. A2 for details).

methods [NL07], in particular degree-corrected SBM, because of their weaker assumptions on the possible structure of ω.

The results summarized in Fig. 3.4, averaging across both directed and undirected networks (Source-Basin structures are averaged over directed networks), show that assortative relationships are the dominant relationship type in most networks for all methods. In the three methods that target such structures (Spectral, Infomap, and Louvain), it is the dominant structure in almost all cases (one exception of Louvain is found for a network with only three communities detected, see Appendix Fig. A.1 for details) and there are in fact very few networks in which non-assortative structures are detected. More interestingly, the SBM method found non-assortative community relationships in a large fraction of the networks (92%). This corresponds not only to the well-studied core-periphery relationship (84%), but

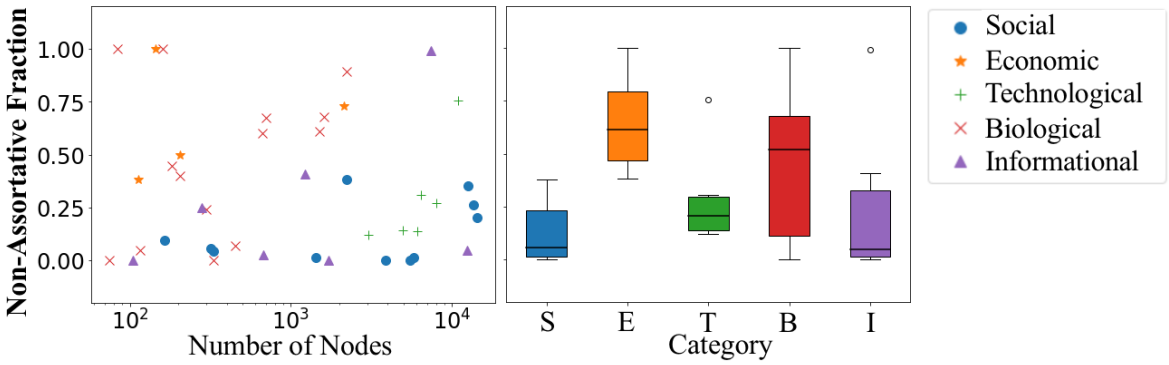


FIGURE 3.5. Non-assortative relationships are found in a variety of empirical directed and undirected networks. Left: fraction of non-assortative relationship (y-axis) as a function of the number of nodes (x-axis), with each symbol representing a network (coloured according to their categories). Right: Boxplot of non-assortative fraction for networks grouped into their categories: Social (S), Economic (E), Technological (T), Biological (B) and Informational (I). The results correspond to the communities obtained using the SBM method.

also the disassortative (54%) and the intriguing source-basin relationship (81% of directed graphs). A diverse distribution is also found using the DNGR method, with an even stronger presence of the source-basin relationshop (92%). Similar results are observed when we restrict the analysis to only undirected or directed networks, or compute the average fraction of different community types, see Appendix Sec. A3. Overall, this finding demonstrates that non-assortative communities are not only a theoretical possibility, they are typical in empirical networks, as long as one uses methods that allow for this possibility and do not focus exclusively on finding assortative partitions. These results confirm also the practical importance of the theoretical advantages of inferential methods such as SBM, in particular of not being restricted to specific community structures.

Given the flexibility of SBM described in Sec. 2.3.4, including the method's freedom to discover an optimal number of communities, we focus on this method and compare results obtained in networks from 5 different domains. Figure 3.5 shows that economic and biological networks have overall higher non-assortative fraction – with median fraction larger than 0.25 and higher variability – while social, technological, and informational networks have fewer non-assortative interactions. In particular, the fact that SBM found such arrangements indicates that they provide a more plausible partition of the nodes into communities than any

partition that does not involve them. In the next section we obtain further insights on the interpretation and significance of source-basin structures through the detailed study of two particular cases.

3.4 Case Studies

We consider two case studies that aim to understand the significance and interpretation of the less studied, non-assortative, community relationships that we found to be common in the previous section. We focus on the SBM method because of its ability to detect such patterns. We use a degree-corrected SBM [KN11] so that the community allocation of nodes is not dominated by their number of links (degree) but instead by how these links are distributed across other communities. We increase the number of blocks B from two until we detect the first non-assortative communities because we are interested in the simplest settings in which they appear.

3.4.1 Political Blog Network

The first case we consider is the famous blog network introduced in Ref. [AG05]. Each node corresponds to a political blog active during the U.S. Presidential Election of 2004 and each directed edge $j \mapsto i$ corresponds to a hyperlink from blog i to blog j (this choice of edge direction corresponds to our convention of aligning it to influence or information flow). Nodes have been self-reported or manually tagged as 'liberal' and 'conservatives', depending on the alignment to the two main parties in USA's political election in 2004. This network has been extensively studied, e.g., the retrieval of the two groups has often been used as a benchmark in the development of new community-detection methods. Importantly, only assortative and core-periphery communities have been reported [KM17; Pei19], in line with the limitations of traditional community-detection methods which target exclusively these types of structures. In Fig. 3.6(a) we show that our more general approach finds source-basin structures as a statistically significant partition of nodes already for $B = 5$. This analysis reveals a fundamental difference between the liberal and conservative blogosphere, with

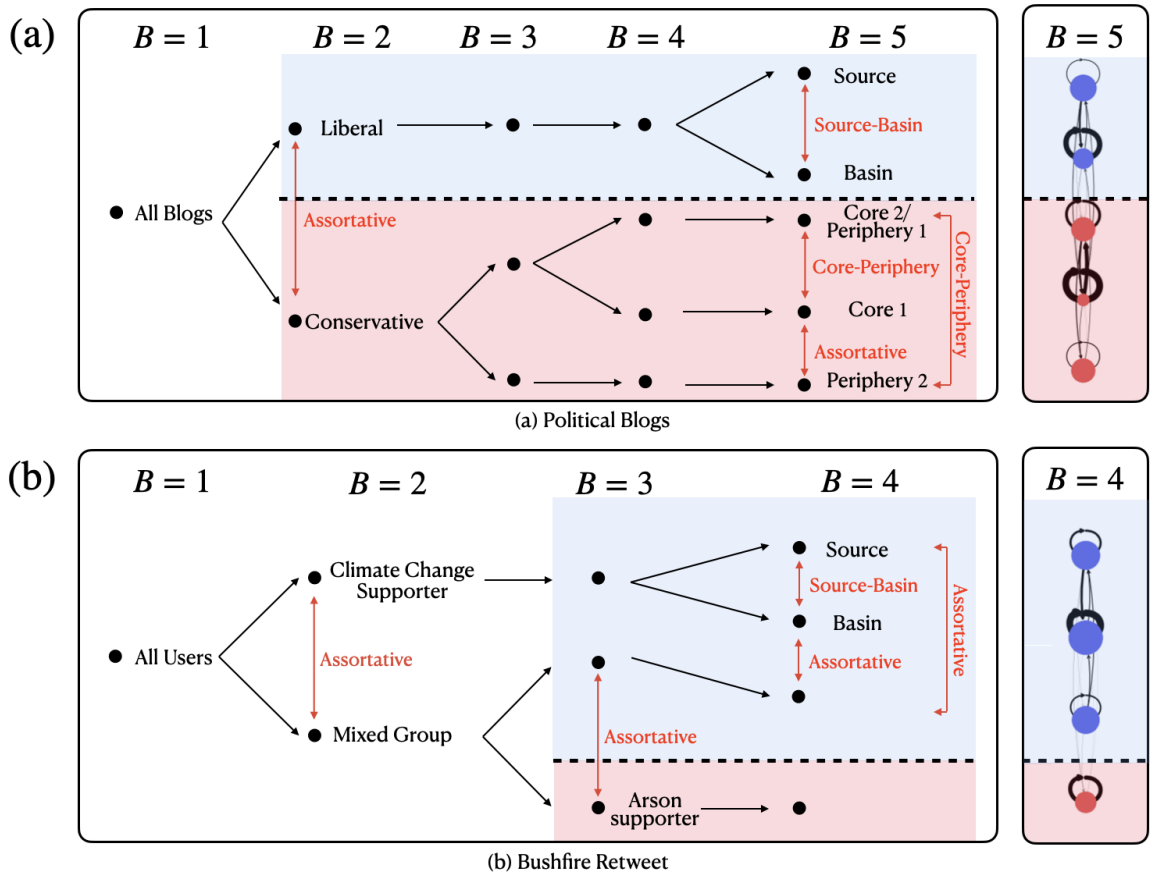


FIGURE 3.6. New types of community structures are typical in online social networks. Affiliation separation in (a) the political blog network [AG05] and (b) the bushfire retweet network [Bed+22]. The partition into groups was obtained using the degree-corrected SBM method with increasing number of communities B. The main plot shows schematically how communities divide and the type of pairwise interaction they share (communities at larger B typically have the same interactions as their parent communities at lower B so that only new interactions are indicated). The plot at the right shows for the largest B the graphical representation of the density matrix ω (see Appendix Fig. A.2 for the complete classification). The colours and labels in the plots (starting at $B = 2$) are based on the manual classification of the political stance of users (in all labeled and coloured groups, more than 73% of the classified users shared the same classification).

conservative blogs organized internally in assortative and core-periphery structures, while liberal blogs show the unusual source-basin structure. Looking at the 10 most referenced blogs (top in out-degree), 4 belong to the source community and none to the basin community. This surprising finding for the liberal blogosphere suggests that the information-source blogs (being

heavily referenced) are not tightly connected with each other while information-receiver blogs are less referenced by others but highly active in referencing others and strongly connected with each other.

3.4.2 Bushfire Network in X

Having shown the appearance of new community structures in a well-known network, we now focus on a social-media example obtained following practices in computational social science (we explain in Sec. 5.2). The nodes in the resulting network correspond to 4,029 X users that were influential in the political debate around the role of climate change in the severe bushfires in Australia in 2019-2020. The directed links $i \mapsto j$ reflect user j retweeting a message originating with user i (information flows from i to j). A fraction of the most influential messages were manually tagged as supporting or rejecting (i) a link between the bushfires and climate; or (ii) a link between the bushfires and the action of arsonists [Bed+22]. This allows us to identify users that support the causal connection between climate change and the bushfires and those that deny it in favour of the (unsubstantiated) claim that the fires were due to arsonists. Our community detection results shown in Fig. 3.6(b) confirm that the separation of these users into assortative communities is a dominant feature of the network, and corroborates the significance of our approach (which is not based on content). Beyond this expected result, we find that already with $B = 4$ communities there are non-assortative relationships within the major assortative separation. Looking at the users in these communities, we find that the source community contains high-profile users (9 of the top-10 most retweeted users, 31% of the users are verified) who retweet more rarely (average in-degree is 99). In contrast, the basin community has fewer high-profile users (0 of the top-10 most retweeted users, only 7% of the users are verified) and they are more active in the debate (average in-degree is 225).

Our finding and analysis of the source-basin structures in both cases discussed above suggests a new picture for the flow of information in online networks. The group that act as a source of information (within one of the sides of the polarized political debate) contains the more central (high-status) nodes which do not interact significantly with each other. This possibly

happens because the source-users compete for the attention of the basin-users within their side of the political division (i.e., on the same side of the main assortative division of the network). In the basin community, the receivers are more strongly connected with each other, discussing and referencing both source nodes and each other during the debate.

3.5 Discussion

This chapter introduced a methodology to identify and quantify the prevalence of unusual communities in networks that can be applied to any directed multi-graphs. While the majority of identification of mesoscale structures assume communities to have either assortative or core-periphery connectivity patterns, we showed that this does not exhaust the possible connectivity patterns between communities and that these additional patterns are not only possible, they are typical – as long as communities are detected using methods that are agnostic about the underlying type of communities – and essential to understand the underlying complex system.

In particular, we find a new structure, source-basin, that reveals a new type of organization of social-media networks that has evaded previous analysis (mostly based on methods that target specific structures) and is also typical in other networks. The appearance of source-basin in both a paradigmatic network and a new social-media network confirms the potential of this structure to reveal new insights into the organization and function of networks. To gain a better understanding of the formation of these structures, a complimentary approach to our data-driven methods of generative models is introduced in next chapter, which give rise to the different community types we detect.

CHAPTER 4

Generative Model

4.1 Motivation

In Chapter 3, we introduced a classification of the possible pair-wise structures in the undirected and directed networks into four types: assortative (A), core-periphery (CP), disassortative (D), and source-basin (SB). In this chapter, we seek for a simple network evolution model that can generate all these structures.

Simple network-evolution models have been a fundamental tool for a deeper understanding of the origin of network properties [New03; GS09]. Network growth models producing community structures which are assortative [Sma+08] and disassortative [KE02] have been introduced, providing insights into the drivers of these community types, while core-periphery community structures have been found to emerge when a small amount of assortative attachment occurs in the presence of preferential attachment [Ure+23]. Rewiring processes applied to existing networks have shown how assortative community structures may emerge from random networks [XS04]. More broadly, the drivers of segregation have been uncovered in Schelling-type network models [HPZ11]. While most studies focused on undirected networks, models with edge direction that combine preferential attachment and homophily have been proposed and used to study the inequality in network algorithms [Esp+22].

In line with this tradition, in our work, we introduce a simple model that can explain the emergence of all four community structure types observed in directed networks. Our model considers generic processes that change the connectivity of nodes – adding, removing, or

rewiring edges – and we observe the structure types that emerge as the long-term consequence of this dynamical process of network evolution.

4.2 Model

4.2.1 General setting

Our goal is to find a simple network-evolution model that reproduces all four types of community-structure discussed in Chapter 3. Accordingly, we consider the following simplifying assumptions:

(i) the number of nodes is fixed, $N(t) = N =$ constant, and we focus on the evolution of the network connectivity $A_{ij}(t)$ over time t.

(ii) the allocation of each node in one of the two groups is known and remains unchanged over t. We denote by $g(i)$ the group $g \in \{0, 1\}$ of node i. The sizes of each group (the number of nodes belonging to them) is given by $N_g = \sum_{i=1}^{N} \delta(g(i) - g)$, where $\delta(x) = 1$ if $x = 0$ and $\delta(x) = 0$ for $x \neq 0$.

(iii) nodes in the same group g are identical (i.e., follow the same evolution rules).

(iv) nodes have agency in determining their incoming edges but not their outgoing ones (i.e., they can deliberately choose who they are influenced by but not who they influence).

We are interested in the equilibrium properties ($t \to \infty$) of large ($N \to \infty$) sparse ($\langle A_{ij} \rangle \to 0$) networks. The choice of the initial network and group allocation is essential to determine the sizes N_g of each group, as discussed above, but otherwise it will play no significant role in the final state of the network or the community-relationship between the groups (except in particular cases in which $\langle z \rangle$ remains constant). In our simulations we start with an Erdős-Rényi directed graph with N nodes, $\frac{\langle z \rangle}{N-1}$ edge creation probability, and random group allocation of nodes.

4.2.2 Network evolution

We will evolve the connectivity of the network over time t step by step i.e. $A_{ij}(t) \mapsto A_{ij}(t + \Delta t)$; compute the density matrix in Eq. (3.2) at each time iteration $t \mapsto t + \Delta t$; and focus on the community-type classification defined in Eq. (3.4). To evolve $A(t)$, we randomly pick a *focal* node i and decide at random between two types of modification of its incoming edges A_{ji} (moves): a swap (with probability P^S) or a change in number (with probability $1 - P^S$) defined as follows:

Swap move: In this move we intend to model the extent into which nodes prefer assortative edges (i.e., edges from nodes in the same group). This is implemented by randomly picking two edges

- one existing (incoming) edge $k \mapsto i$ of node i (such that $A_{ki} = 1$); and
- a non-edge $j \mapsto i$ (i.e., $A_{ji} = 0$) as a candidate edge,

and deciding whether we rewire from edge $k \mapsto i$ to edge $j \mapsto i$ (at time $t + \Delta t$) depending on the consistency between their groups $g(k), g(j)$, and $g(i)$. We give a preference to an assortative edge (i.e., one coming from a node in group equal to $g(i)$) with probability $P^A = P^A_g = P^A(g(i))$, in which case we keep the edge coming from a node with group equal to $g(i)$. This implies that with probability $1 - P^A$ a preference is given to disassortative edges, in which case we keep the edge coming from a node with group different from $g(i)$. If both candidate edge $j \mapsto i$ and existing edge $k \mapsto i$ are such that $g(j) = g(k)$, we randomly pick one to keep and delete the other. We investigate the simplest case in which the two edges are chosen randomly (from all existing and non-existing edges). Alternative choices could consider local searches.

Change move: In this move we intend to model the process of adding and removing edges, and therefore how selective nodes are about the nodes they are influenced by. We first decide either to remove or add edges with probability $\frac{1}{2}$ (as shown in Appendix B2, this choice is taken without loss of generality). To add an edge, we pick a random node j that does not

have an edge pointing to i (i.e., $A_{ji} = 0$) and add this edge $j \mapsto i$. To remove, we delete with probability $\alpha = \alpha_g = \alpha(g(i))$ each of the incoming edges of node i.

A summary of the evolution is given in Fig. 4.1 and a list of parameters in Table 4.1. The implementation of our model can be found in the repository [Liu24]. The choice and application of each move can depend on the group of the involved nodes $g(i)$, but it remains the same for all t. We consider $\Delta t = 1/N$ so that $\Delta t \to 0$ as $N \to \infty$ and $t \mapsto t + 1$ after N steps (i.e., on average each node is picked once).

Parameter	**Description**	**Choice**
	Primitive Parameters	
P_g^S	Probability of swap move	$P_0^S = P_1^S = P^S$
P_g^A	Probability of assortative move	$0 \leq P_g^A \leq 1$
α_g	Probability of removing in-edge	$0 \leq \alpha_g \ll 1$
N_g	Number of nodes	$N_g \gg 1$
	Derived Parameters	
β_g	Proportion of in-edges coming from same group among all in-edges	Eqs. (4.3) and (4.6)
$\langle z \rangle_g$	Average in-degree	$\langle z \rangle_g = 1/\alpha_g$, Eq. (4.2)
b	Ratio between average in-degrees	$b \equiv \langle z \rangle_0 / \langle z \rangle_1$
c	Ratio of group sizes	$c \equiv N_0/N_1$

TABLE 4.1. List of parameters used in our model (4 primitive parameters) and calculations (4 derived parameters). In general, each parameter depends also on the group g of the focal node ($g \in \{0, 1\}$). The last column indicates the particular choices and ranges.

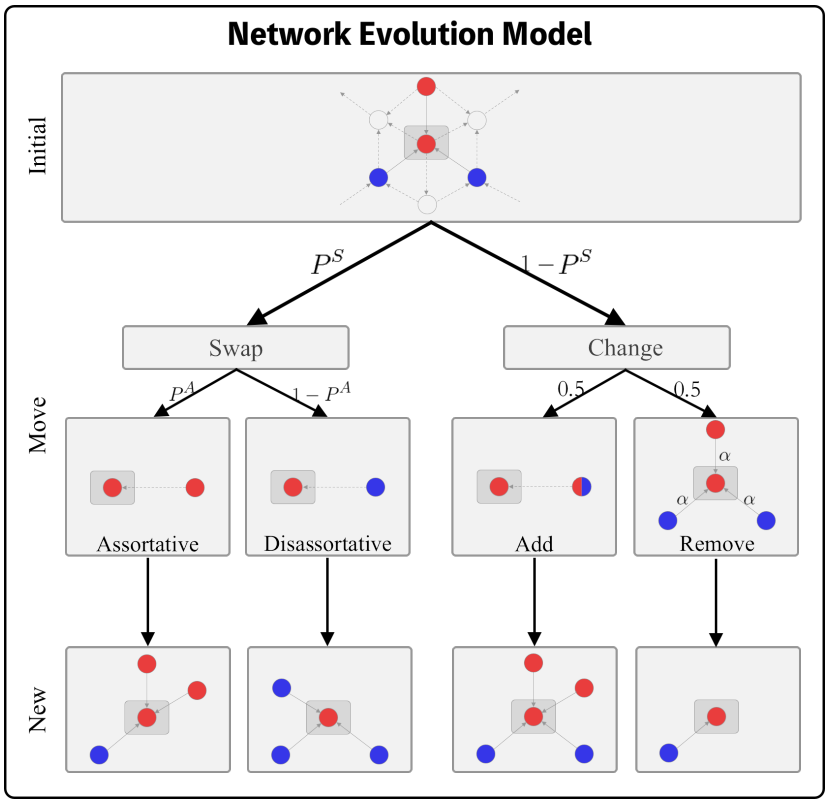


FIGURE 4.1. The network evolution model represented as a decision tree. The picture depicts the possible one-step alterations in the connectivity of the network around the randomly chosen focal node (highlighted in the centre of the top panel). The groups of nodes are represented in the picture by colours $g(i) = 0$ in red and $g(i) = 1$ in blue. The swap move allows users to give preference to edges (influence) from a user of the same group (assortative choice with probability P^A) or from the different group (disassortative choice with probability $1 - P^A$), while the change move controls the number of incoming edges users in a group will have on average (controlled by parameter α). The algorithmic implementation of this model is described in Appendix Algorithm 3 and codes can be found in the repository in Ref. [Liu24].

4.3 Structure types obtained from the model

Here we compute the types of structure – A, D, CP, SB as defined in Fig. 3.2 – obtained for $t \to \infty$ for different cases of the model introduced above. This is obtained by computing the long-time $t \to \infty$ density matrix ω in Eq. (3.2) for the different parameters of our model $P_g^S, P_g^A, \alpha_g, N_g$ and checking if the inequalities (3.4) defining each community type are

satisfied. We are interested in finding the necessary conditions for the appearance of each type and the parameter regime for which a type appears.

Our calculations rely on a key property of our model: the swap move controls assortativity (while the change move is neutral about it) and the change move controls the in-degree (while the swap move preserves it). Accordingly, we start our analysis by computing the average in-degrees $\langle z\rangle_0$ and $\langle z\rangle_1$ as a function of the parameters of the change move and then proceed with the analysis of the effect of the swap move at fixed average in degree.

4.3.1 Change move determines the average in-degree

The primary difference between the remove and add moves is that the former acts on each edge while the latter is a single addition. For large average in-degrees $\langle z\rangle$ and small $\alpha \ll 1$, we can consider the action of this move as providing only a small modification of the in-degree in a single time increment $dt \sim \Delta t = 1/N$. In a continuous and mean-field approximation, the average degree $\langle z\rangle$ of nodes in any of the groups g will change as

$$\frac{d\langle z\rangle}{dt} = -\alpha\langle z\rangle\frac{1}{2} + \frac{1}{2}, \tag{4.1}$$

and therefore converge for large t and N towards the fixed point

$$\frac{d\langle z\rangle}{dt} = 0 \Leftrightarrow \langle z\rangle = \langle z^*\rangle = \frac{1}{\alpha}, \tag{4.2}$$

where for simplicity we omit the dependence on g.

4.3.2 General Solution

In this section we will compute the expected density matrix ω_{rs} (at $t \to \infty$) as a function of the model parameters. We define $\beta_r(i)$ to be the probability that one of the in-edges of node i is coming from a node j in the same group $g(j) = g(i) = r$. Denoting by e_{sr} the total number of edges from nodes in group s to nodes in group r, β_r at time t can be computed in a mean-field approximation (i.e., for all i with the same $g(i) = r$) by the macroscopic

assortativity of group $g(i) = r$ as

$$\beta_r(t) = \frac{e_{rr}(t)}{e_{rr}(t) + e_{sr}(t)} \tag{4.3}$$

with $s \neq r$. The denominator of Eq. (4.3) is the total incoming edges for nodes in group r. We obtain the one-time evolution of $\beta_r(t)$ by considering how the number of within-group edges e_{rr} changes in one time step. This is done by computing the probabilities $P^{\pm}$ of increasing (+) and decreasing (-) edges for each of the moves (swap or change) and incorporating the effects into e_{rr}

$$e_{rr}(t + \Delta t) = e_{rr}(t) + P^{\text{swap}+} - P^{\text{swap}-} + P^{\text{change}+} - P^{\text{change}-}\langle z \rangle_r, \tag{4.4}$$

where the last term is multiplied by the average degree $\langle z \rangle$ because the removal of edges in the change move (with probability α) applies to each of the existing edges independently. Expressing the probabilities $P^{\pm}$ in Eq. 4.4 as a function of the model parameters, and using Eq. (4.3), we obtain (see Appendix B1)

$$e_{rr}(t + \Delta t) = Be_{rr}(t) + C, \tag{4.5}$$

with $B = B(t) = 1 - P_r^S \frac{N_r}{N(e_{rr}(t)+e_{sr}(t))}[P_r^A \frac{N_r}{N} + (1 - P_r^A)(1 - \frac{N_r}{N})] - (1 - P_r^S)\frac{N_r}{2N(e_{rr}(t)+e_{sr}(t))}$ and $C = C(t) = [P_r^S P_r^A + \frac{1}{2}(1 - P_r^S)]\frac{N_r^2}{N^2}$. From Eq. (4.5), we obtain the solution as a function of the initial condition at time $t = 0$ as

$$e_{rr}(t) = B^{Nt} e_{rr}(t = 0) + \frac{B^{Nt} - 1}{B - 1} C.$$

Since $B < 1$ for all t, $B^{Nt} \to 0$ and $e_{rr}(t) \to \frac{C}{1-B}$ for $t \to \infty$. In this limit (system at equilibrium), we assume that the total number of edges $e_{rr}(t) + e_{sr}(t)$ is a constant (the expected value at equilibrium) and that the average in-degree can be computed by Eq. (4.2). Taking this into account, and introducing the results derived from Eq. (4.5) in Eq. (4.3), we obtain the equilibrium probability of an assortative in-edge as

$$\beta_r = \frac{P_r^S P_r^A + \frac{1}{2}(1 - P_r^S)}{P_r^S(P_r^A + (1 - P_r^A)\frac{N_s}{N_r}) + (1 - P_r^S)\frac{N}{2N_r}}. \tag{4.6}$$

We can now use Eq. (4.6) to compute the expected number of edges between groups and therefore the density matrix ω_{rs}. Since in our model we only have two groups, an edge can only be from one group or the other. The probability that an in-edge of node i is from a different group is thus $1 - \beta_r$. The number of in-edges from nodes in the same group is thus $z_i\beta_r$ and the number of in-edges from the different group is $z_i(1 - \beta_r)$. Taking this into account, we can compute the number of within group edges e_{rr} and between group edges e_{sr} as the sum of all edges $\ell \equiv j \mapsto i$ as

$$e_{rr} = \sum_{i=1}^{N} \delta(g(i) - r) z_i \beta_r = N_r \langle z \rangle_r \beta_r \tag{4.7}$$

$$e_{sr} = \sum_{i=1}^{N} \delta(g(i) - r) z_i (1 - \beta_r) = N_r \langle z \rangle_r (1 - \beta_r) \tag{4.8}$$

The density matrix – defined in Eq. (3.2) – of the network with two groups is then computed as

$$\omega = \begin{bmatrix} \frac{e_{00}}{N_0(N_0-1)} & \frac{e_{01}}{N_0 N_1} \\ \frac{e_{10}}{N_0 N_1} & \frac{e_{11}}{N_1(N_1-1)} \end{bmatrix} = \begin{bmatrix} \frac{\langle z \rangle_0 \beta_0}{N_0-1} & \frac{\langle z \rangle_1 (1-\beta_1)}{N_0} \\ \frac{\langle z \rangle_0 (1-\beta_0)}{N_1} & \frac{\langle z \rangle_1 \beta_1}{N_1-1} \end{bmatrix} \approx \begin{bmatrix} \frac{\langle z \rangle_0 \beta_0}{N_0} & \frac{\langle z \rangle_1 (1-\beta_1)}{N_0} \\ \frac{\langle z \rangle_0 (1-\beta_0)}{N_1} & \frac{\langle z \rangle_1 \beta_1}{N_1} \end{bmatrix} \tag{4.9}$$

where β_0 and β_1 are given by Eq. (4.6), $\langle z \rangle_0$ and $\langle z \rangle_1$ by Eq. (4.2), and the approximate expression is used for large networks ($N_0, N_1 \gg 1$ so that we can drop the subtraction of 1 in the denominator).

The density matrix in Eq. (4.9), evaluated using Eq. (4.6), is the main result of our calculations. In Fig. 4.2 we show how the results of a direct simulation of our model compares to the prediction from these equations, confirming the agreement of the average density at long times. Next we discuss the different types of community relationships – A, CP, D, and SB, obtained from ω from Eq. (3.4) – for different combinations of the model parameters – $N_0, N_1, P_0^A, P_1^A, P_0^S, P_1^S, \langle z \rangle_0$, and $\langle z \rangle_1$ – that appear in the computation of the density matrix in Eq. (4.9).

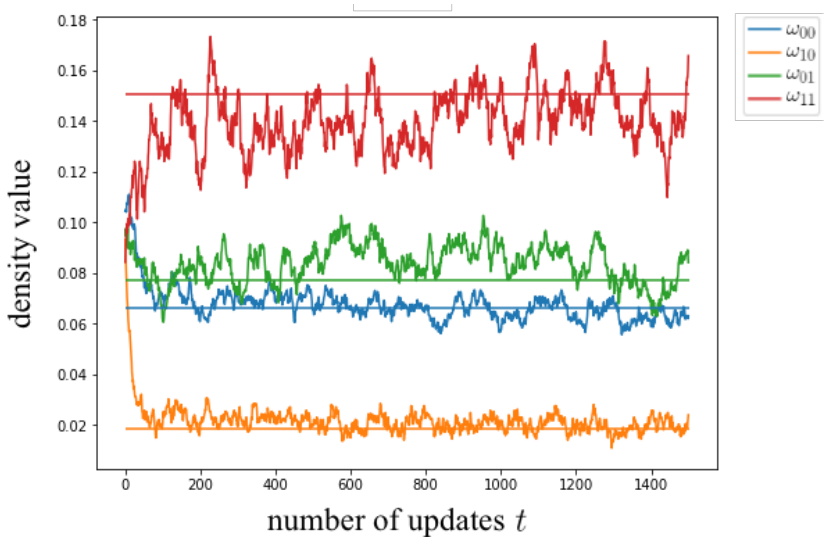


FIGURE 4.2. Density of edges ω_{rs} between nodes in groups 0 and 1 as a function of time t. The curves were obtained performing a direct simulation of our model with parameters $P_0^S = 0.7, P_1^S = 0.5, P_0^A = 0.9, P_1^A = 0.8, \alpha_0 = 0.2, \alpha_1 = 0.1, N_0 = 67, N_1 = 33$. The horizontal straight line correspond to the expected value for $t \to \infty$ and $N \to \infty$ obtained from Eq. (4.9).

4.3.3 Particular Cases

Equal average degrees $\langle z \rangle_0 = \langle z \rangle_1 \Rightarrow$ no SB.

From the definition (3.4), the following conditions both hold if SB structure with group r as basin and group s as source is observed

$$\omega_{rr} > \omega_{rs} \Rightarrow \frac{e_{rr}}{N_r - 1} > \frac{e_{rs}}{N_s} \approx \frac{e_{rr}}{N_r} > \frac{e_{rs}}{N_s} \tag{4.10}$$

$$\omega_{sr} > \omega_{ss} \Rightarrow \frac{e_{sr}}{N_r} > \frac{e_{ss}}{N_s - 1} \approx \frac{e_{sr}}{N_r} > \frac{e_{ss}}{N_s} \tag{4.11}$$

with $r, s \in \{0, 1\}$, $r \neq s$ and large N_r and N_s. These two inequalities imply that $\frac{e_{rr}+e_{sr}}{N_r} > \frac{e_{ss}+e_{rs}}{N_s} \Rightarrow \langle z \rangle_r > \langle z \rangle_s$. This is in contradiction to our hypothesis of $\langle z \rangle_r = \langle z \rangle_s \Rightarrow \frac{e_{rr}+e_{sr}}{N_r} = \frac{e_{rs}+e_{ss}}{N_s}$. This proof by contradiction, which applies not only to our model but to any network partition into two groups, implies that a difference in average degree is essential for the existence of the SB structure (the basin requires a higher in-degree than the source). This proof by contradiction implies that a difference in average in-degree is a necessary condition for the existance of the SB structure. The basin requires a higher in-degree than the source, in line with the view that the the basin actively receives influence from both source and itself.The

reversal of link direction maps in-degree to out-degree and SB to CP, so that equivalent computations imply, by symmetry, that an *out*-degree difference between the two groups is a necessary condition for the existance of CP structures. These two conditions apply not only to our model but to any network partition into two groups.

Equal assortativity $P_0^A = P_1^A \Rightarrow$ no CP.

The appearance of CP structure with group $r \in \{0, 1\}$ as core and group $s \in \{0, 1\}$ ($s \neq r$) as periphery implies the following two mutually conflicting inequalities

$$\omega_{rr} > \omega_{sr} \Rightarrow \beta_r > \frac{N_r}{N} \Rightarrow P_r^A > \frac{1}{2}, \tag{4.12}$$

$$\omega_{rs} > \omega_{ss} \Rightarrow \beta_s < \frac{N_s}{N} \Rightarrow P_s^A < \frac{1}{2}, \tag{4.13}$$

where in the first step we used Eq. (4.9) and in the second step we used Eq. (4.6). This demonstration shows that groups with equal tendency to establish assortative edges $P_0^A = P_1^A$ could not support a CP structure and the CP structure only appears when one group prefers assortative edges and the other group prefers disassortative edges.

Swapping regime $P^S \to 1_-$.

Here we consider the case in which swap moves dominate over change moves. We denote this by $P^S \to 1_-$ to convey that there are enough change moves to allow the average degree to converge in Eq. (4.2) but that otherwise it plays no role. The probability of linking to the same group, β_r, is obtained by setting $P^S = 1$ in Eq. (4.6) leading to

$$\beta_r = \frac{P_r^A}{P_r^A + (1 - P_r^A)\frac{N_s}{N_r}}. \tag{4.14}$$

Using this expression in the computation of the density ω in Eq. (4.9) we obtain a simpler separation of the parameter space into the community types. Fig. 4.3 shows that the four structure types appear in the parameter space of P_0^A and P_1^A. An A relationship appears when both groups strongly prefer linking to nodes in the same group, while D is formed in the opposite situation. The CP structure appears when one group prefers assortative linking (being the core) while the other group (the periphery) prefers to link to the different group, i.e. it is influenced more by the core group than its own group. SB occurs for intermediate values of

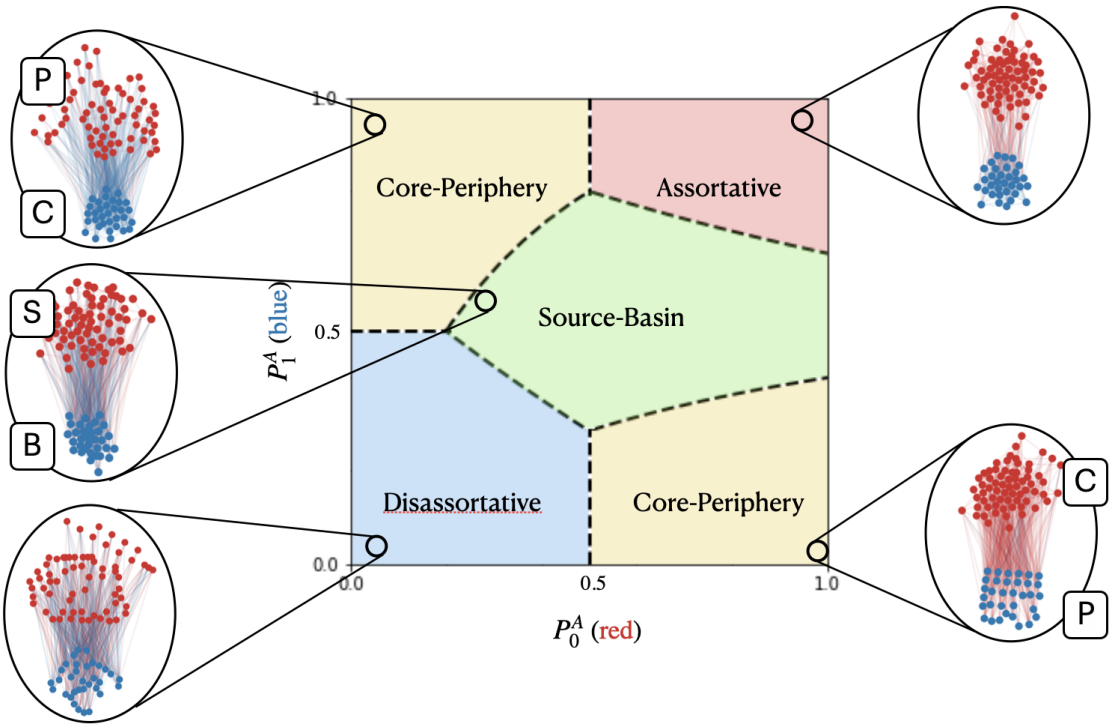


FIGURE 4.3. Parameter space for Eq. (4.9), which is valid in the limit $P^S \to 1_-$ with swap and change moves. The group size ratio $c = \frac{N_0}{N_1} = 2$ and the average in-degree ratio $b = \frac{\langle z \rangle_0}{\langle z \rangle_1} = 0.5$. Insets: examples of networks obtained at the indicated parameters through direct numerical simulations of our model, plotted using the NetworkX package. The node color indicates the group membership of the node and the edge color indicates information source (same as source node color).

P^A, most strongly around $P_0^A = P_1^A = 0.5$ (no preference in terms of groups). Fig. 4.4 shows the effect of asymmetric group sizes ($N_0 \neq N_1$) and average degrees ($\langle z \rangle_0 \neq \langle z \rangle_1$), indicating that the SB region in parameter space grows and is deformed with increasing asymmetry.

The effect of change moves $P^S < 1$.

We now consider the effect of change moves in the picture discussed above (i.e., when $P^S < 1$). For simplicity, we assume the two groups have the same probability of choosing the swap move, i.e. $P_0^S = P_1^S$. The effect of P^S in reproducing the four community types is shown in Fig. 4.5. As less swap moves are performed (P^S decreases), more in-edges

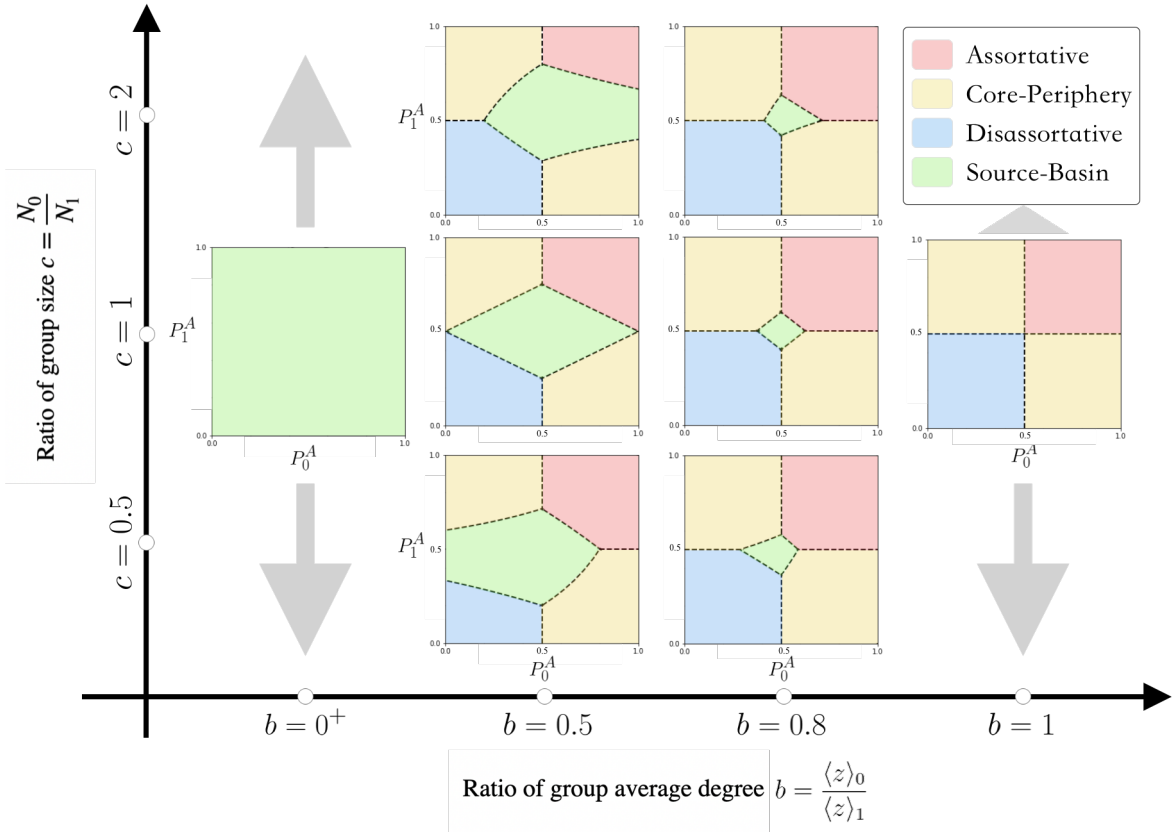


FIGURE 4.4. Effect of group-asymmetry in group size N (y axis) and in-degree $\langle z \rangle$ (x axis) in the parameter space of assortative linking shown in Fig. (4.3). The boundary of all structure types is computed from Eqs. (4.9), (4.14), and (3.4) in the limit of $t \to \infty$, $N \to \infty$, and $P^S \to 1\_$.

among all edges come from the change move and therefore the assortative probability P^A of the swap move plays a less significant role. The degree difference between the groups becomes increasingly the only important distinction between them. As a result, the area of SB in the P^A parameter space increases with reducing P^S. Surprisingly, there is a finite value $P^{S*} > 0$ such that SB is the only community type for any $P^S < P^{S*}$ (see Appendix B3 for a computation of P^{S*}). An intuitive explanation for these results is that the change move does not take the groups (colours) of nodes into account and therefore it effectively leads to a weakening of the assortative moves (P^A comes closer to 0.5), which is the SB region in Fig 4.3.

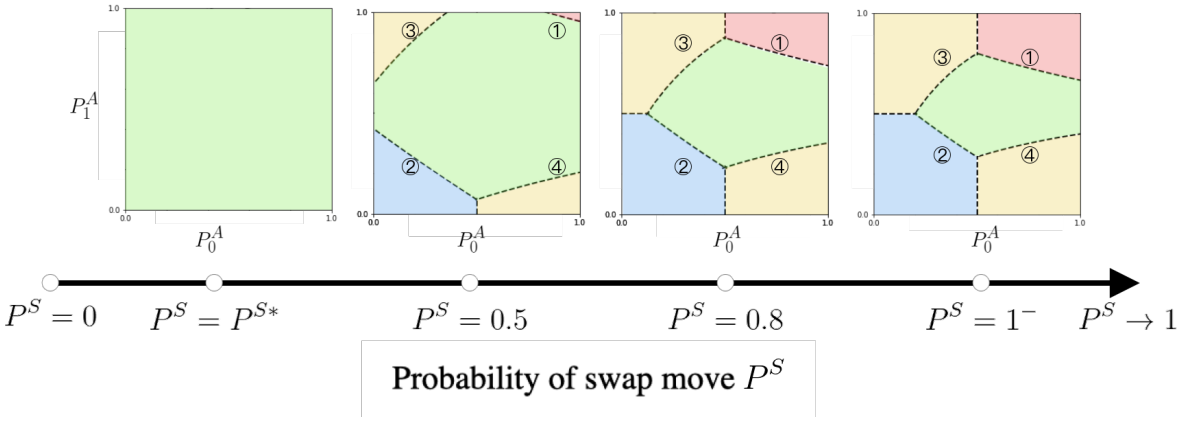


FIGURE 4.5. Probability of swapping P^S affects the conditions for the appearance of different community relationships. The results shown in the figure were computed using $c = \frac{N_0}{N_1} = 2$, $b = \frac{\langle z \rangle_0^*}{\langle z \rangle_1^*} = \frac{1}{2}$, and $P_0^S = P_1^S$. The boundary of all structure types is computed with Eqs. (4.9), (4.6), and (3.4), valid in the limit of $t \to \infty$ and $N \to \infty$. Lines 1-4 indicate the boundaries between A,D,CP and SB, see details in Appendix B3.

4.4 Discussions

In this chapter, we proposed a simple network-evolution model that generates directed networks with the four different types of relationships between 2-communities of nodes. Our mean-field calculations of the dependence of the relationship type (A, CP, D, SB) on the model parameters reveal the necessary conditions and possible mechanisms behind their creation. In particular, we find that a CP relationship cannot exist if both groups have the same tendency to establish links within their own communities and that an SB relationship cannot exist if the communities have the same average degree. Our analysis revealed a close relationship between the difference in in-degree between the communities and the appearance of the Source-Basin community type. These results generalize to the case of directed networks to previous analyses on core-periphery in undirected networks, which have shown that CP structure between two groups is highly related with degree difference [BL16; ZMN15] and that no CP structure is found in two groups using a degree-correction method [KM18]. To further clarify the connection between degree and structure type, we consider an alternative definition of a density matrix $\tilde{\omega}$ in which the number of edges e_{rs} is normalized by the product of existing edges $e_r e_s$ instead of the possible edges $N_r N_S$. This choice corresponds to relative

connectivity between communities, normalized by their propensity of links, and is similar to the approach used in degree-corrected stochastic block models. In this case, neither CP nor SB are possible (as shown in Appendix Sec. B4) in the case of two communities, consistent with previous observations (only A and D are possible).

CHAPTER 5

Social Media Data

5.1 Introduction

In this chapter we apply the methods and ideas discussed in the previous chapters to the study of social media. Social media platforms, including X (former Twitter), Facebook, YouTube, Instagram and LinkedIn, have become prevalent for individuals to express their opinions, access news and information, and engage with others anytime and anywhere. The vast amount of information from social media provides opportunities for the analysis of various problems such as political polarization [Cin+21; CM15], information diffusion [Gle+14] and opinion mining [Mit+13; Cod+15]. In particular, identifying groups of users with similar connectivity pattern in a network approach attracts increasing attention in sociology studies, and community detection becomes a standard practice that has been applied in many areas as described in Sec. 2.4.

While previous analysis of mesoscale structures focuses on either assortative or core-periphery connectivity patterns, our methodologies developed in Chapter 3 indicate that other patterns are also possible and typical in networks, and can also reveal new types of community structure [LAA23]. It motivates us to consider this approach in social media networks to gain a deeper understanding of the organization and function of networks, taking advantage of the interpretability of social media data. In this chapter, we propose a quantitative analysis of social media data on X, focusing on interactions between users and textual data of tweets. We measure the public mood from textual data of tweets and find communities based on user activity including retweets and replies.

5.2 X Data

We focus on the X platform due to its easy access to information including user's profiles and tweets. X allows users to communicate with friends or strangers by retweeting, replying, or mentioning in tweets, which is a short textual message that can include multimedia content. In our analysis, we include three datasets about specific events on X.

The first dataset is about Australian bushfire discussion on X from 2019 to 2020. The intensive bushfires drew much attention all over the world and brought lots of discussions in online social media. Specifically, we noticed that a political debate around the role of climate change in bushfires happened on X. The data was obtained using the X streaming API by the co-supervisor A/Prof. Tristram Alexander as follows

(1) Collect all tweets containing the word 'climate' between November 2019 to June 2020, and coming from accounts with Australia-specific locations in the user location metadata.
(2) Identify the top 100 most retweeted tweets each day during this period.
(3) Capture the user timelines for the authors of these daily top 100 most retweeted tweets.
(4) Keep activities in these top-user timelines that retweet or reply to other top users.

The dataset includes 4,071 unique users, 3,184,690 retweets, and 2,158,374 replies between these users. User profile data is available such as verified and followers count, friend count, and status count(see details in Appendix Sec. C3). The classification of these users was obtained from Ref. [Bed+22], where a fraction of users were classified based on their tweets where two categorical classifications were given as follows

- **Arson-Fire** indicates whether the tweet supported the view that bushfires were caused by arsonists.
- **Climate-Fire** indicates whether the tweet supported the opinion that bushfires were related with climate change.

Users were classified into two categories: **supporter** and **opposer** in the Arson-Fire and Climate-Fire classifications according to the majority labels of their tweets. 56 users were selected in Arson-Fire classification and 194 users were selected in Climate-Fire classification. The textual data of tweets extracted in step (1) was also recorded and split into daily sets. In total, we collected 590,607 tweets and 27 words per tweet on average during 217 days from November 2019 to June 2020.

In our analysis, we include two more X datasets which were obtained from Ref. [Gai+21] about two political events in Germany, that is, Saxon state election voting in September 2019 and a violent attack on police on the New Year's Eve of 2019(NYE) four months later. The election was a nationwide political event where politicians amplified their voices on X. The NYE incident generated a lot of discussion, and a debate about the role of left-wing extremism and the deliberate provocation of police in this violence was raised on X. The authors of Ref. [Gai+21] collected the data based on a set of seed users that were chosen from the election candidates list. All tweets containing seed users in the form of retweets, replies, and mentions were collected. The election datasets were restricted to a subset of tweets containing election-related words from the time period between the 25th of July and the 10th of September. The resulting dataset contains 41,815 users with 129,700 retweets and 80,470 replies. The NYE dataset was restricted to tweets containing incident-specific words from December 31 until January 19. The dataset includes 13,551 with 24,278 retweets and 17,381 replies.

5.3 Characteristic of Bushfire Discussion

With the spontaneous nature of the content shared, social media platforms become a rich resource for evaluating public happiness. By capturing real-time expressions of opinions, emotions, and reactions from millions of users, social media provides valuable insights into societal trends, public happiness, and collective behaviors [Mit+13; Cod+15]. In this section, we measure the public sentiment during the period of the bushfire and identify a series of

events that affect public happiness from the time-series tweets, gaining an understanding of the characteristics of bushfire discussion.

5.3.1 Sentiment Analysis

We compute the sentiment of tweets collected in the bushfire dataset. The sentiment score for a text is calculated based on the individual words used in the text. We use sentiment scores collected by Warriner et al., where 13,915 of the most frequently used English words from four sources were given ratings on valence which indicates the degree of pleasantness. In Ref. [WKB13], each word was rated from 1 *(unhappy)* to 9 *(happy)*. Then the average rating over 20 responses became the word's score. In this project, the valence score of a text is computed by averaging the scores of all words appearing in this text.

Apart from the simple algorithm in Ref. [WKB13], we find that SentiStrength is a powerful tool for informal English texts like tweets [The+10]. SentiStrength also uses a dictionary of words with associated sentiment strength from 1*([no positive/negative emotion]* to 5 *[very strong positive/negative emotion])*. Moreover, Sentistrength developed many algorithms to solve the problem of informal short text such as spelling correction, booster word list, negative word list, etc. From SentiStrength, a single positive/negative scale of a text can be obtained from -5 *(negative)* to +5 *(positive)*.

5.3.2 Result

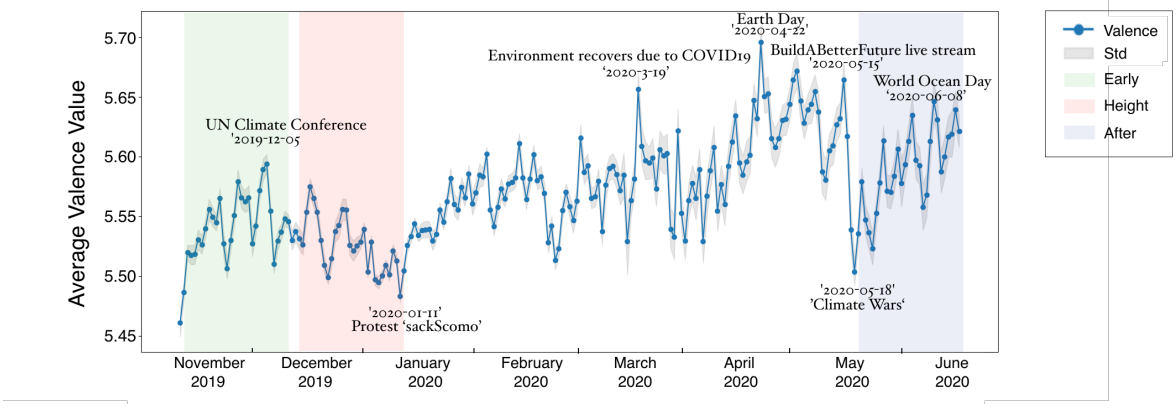
We apply two sentiment methods on the tweets during bushfire discusion. Fig. 5.1 shows the average valence and SentiStrength values of the climate tweets by day during the 8-month time span. The highlighted regions in this plot are early, height, and after stage of bushfire which are defined from 12/11/2019 to 11/12/2019, from 14/12/2019 to 12/1/2020, and from 19/5/2020 to 17/6/2020, respectively [Bed+22]. Fig. 5.1a shows that public happiness is lower than average valence (5.62) in the early and height stages. The figure indicates an increasing trend in the early stage and the peak occurs when the UN climate conference is held. The happiness decreases at the height of the bushfire and reaches its lowest point on the day of the

'sackScomo' protest. After the height stage, the happiness increases slowly, and many climate-supporting events are discussed on X such as "Environment recovered during Covid-19", "Earth Day", "BuildABetterFuture live stream". Public happiness drops dramatically when ABC News publishes a video about the climate policy in Australia called "Climate Wars". After this fall, the happiness rises again and peaks on World Ocean Day. The SentiStrength method in Fig. 5.1b show similar trends with valence which corroborates our finding about the happiness variation during bushfire obtained using valence. The analysis based on daily tweets containing "climate" clearly shows that the public is unhappy about bushfires both in valence and SentiStrength measures. We also find a clear impact of climate-specific events on public happiness that climate-supporting events have a positive effect and bad news has a negative effect.

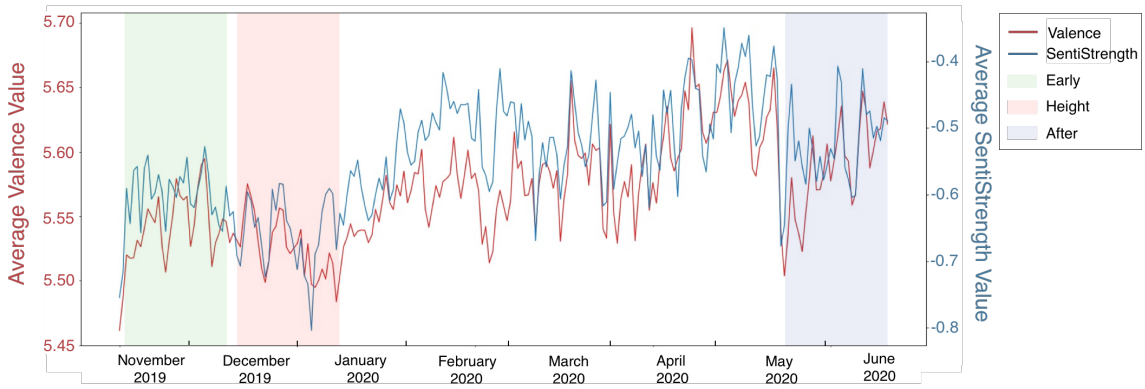
5.3.3 Validation

We validate our machine-learning methods with manual ratings obtained from Ref. [Bed+22] where random 120 tweets per day from selected 15 time periods were manually rated with sentiment *[positive, negative, N/A]*. For simplicity, we assign the sentiment score to be $+1$ for positive, -1 for negative, and 0 for N/A. We compare the automated result with manual ratings by two traditional methods: Pearson correlation coefficient and inter-coder agreement by Cohen's Kappa method [McH12] with a value closer to 1 meaning a stronger correlation. Fig.5.2 shows a strong linear relationship between the automated result and manual rating indicating that our automated methods agree with manual rating. The Pearson correlation coefficient between valence and manual rating is 0.818, and the Pearson correlation coefficient between SentiStrength and manual rating is 0.948 showing that both automated methods perform well with SentiStrength performing closer to the human mind.

We further confirm the agreement with second manual ratings from Ref. [Bed+22] where random 40 tweets per day from the same 15 time periods were rated. In almost all time intervals, we get inter coder reliability $\kappa > 0.81$($\kappa = 0.76$ on 2019-11-24 and $\kappa = 0.64$ on 2020-6-8). Overall, we have $\kappa = 0.88$ showing almost perfect agreement [LK77] between manual ratings and automated results.



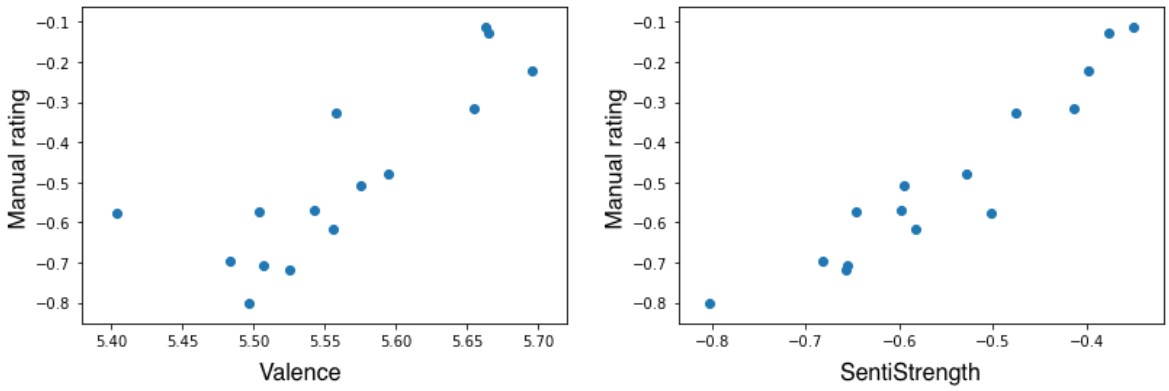
(A)



(B)

FIGURE 5.1. Temporal sentiment trend in X. (a): Average valence of tweets containing the word "climate" from November 2019 to June 2020 by day. The standard deviation in grey is computed by Bootstrapping method where we resample daily tweets with replacement for 1,000 iterations. The green, red and blue regions indicates the early, height, after stage of bushfire defined in Ref. [Bed+22]. (b): Average valence (red) and average SentiStrength (blue) of tweets containing the word "climate" from November 2019 to June 2020 by day. The green, red and blue regions indicate the early, height, after stages of bushfire.

The validation provides further support for our findings of public happiness variation during bushfires. The analysis provides us understanding of this discussion on X that the public has the lowest happiness in the height stage of bushfire and their mood is related to climate-specific events. After the successful studies in textual data of tweets, we are also interested in how online users interact and involved in specific events. In the next section, we find groups of users according to the user's connectivity and analyze the structural pattern within the user's interaction.



(A) (B)

FIGURE 5.2. Sentiment by machine-learning methods (a): valence and (b): SentiStrength against manual ratings. Each point corresponds to a specific date which is randomly selected from time window.

5.4 Community Structure

The identification of community structures based on network topology has been studied in social sciences as described in Sec. 2.4. In this section, we focus on the interaction information between users to have a deeper understanding of user's involvement in political debates. We take retweet and reply interactions of the three events in X since retweet interaction is related to opinion communities while reply interaction reveals user's participation in public debate [Gai+21]. We construct a network with users as nodes and interactions as edges so that we have two types of directed edges, retweet edge and reply edge. The retweet/reply edge is directed from node i to j meaning user j retweet/reply node i, and has edge weight as the multiplicity of retweet/reply. With two types of edges, we construct three networks for each event: retweet-only network, reply-only network, and multilayer network that contains both retweet and reply. The multilayer network as described in Sec. 2.1 is able to include two valuable interaction information in X and provide a richer picture of user's interaction patterns. We focus on the largest connected component of each network with detailed information in Table. 5.1. In the following sections, we apply the methodology developed in Chapter 3 and analyze the community structures, which may reveal insights into political debates on social media platforms. We use degree-corrected hierarchical Stochastic Block Model (hSBM), which is described in Sec. 2.3.4, to detect communities due to its ability of having a larger

resolution limit, providing various levels of fineness of communities, and allowing various degrees within communities. The detailed procedure of community detection can be found in Appendix Sec. C1.

Network	Bushfire		
Interaction	retweet	reply	multilayer
Nodes	4,029	3,954	4,043
Edges	429,235	233,336	
Network	Election		
Interaction	retweet	reply	multilayer
Nodes	29,565	20,238	40,836
Edges	89,603	52,487	
Network	NYE		
Interaction	retweet	reply	multilayer
Nodes	9,232	6,147	13,385
Edges	20,307	12,191	

TABLE 5.1. Network Information for bushfire, election, and NYE datasets. The rows are interaction type of network (Interaction), the number of nodes (Nodes) and the number of edges (Edges).

5.4.1 Bushfire

Hierarchical communities of nodes obtained from the hSBM model based on a directed retweet multi-graph is shown in Fig. 5.3. In the most fine-grained level, level 3, nodes are divided into 9 small communities where each community retweets itself more often than retweets other communities, which corresponds to the assortative structure in classification of Chapter 3. The manual classifications of Arson-Fire and Climate-Fire in Fig. 5.4 show that the supporters of the **climate theory view** that bushfires were related with climate change,

who supports the link of Climate-Fire and deny the link of Arson-Fire, and the supporters of the **arson theory view** that bushfires were caused by arsonists, who hold opposite opinions of two links, are separated into different groups in level 3. It indicates that communities found by the hSBM model are aligned with users' opinions on climate change and convinces us to take a further look at the structural patterns of these communities. We apply the methodology developed in Chapter 3 and find that even though all interactions between communities are assortative (see Appendix Fig. C.2), some communities exhibit a relatively high density of retweeting others. In particular, we notice that two communities with arson theory supporters interact with each other but rarely interact with the other 7 communities that consist of climate theory supporters.

These hierarchical structures provide not only a single representation of user communities but also reveal how small communities merge at various levels. This offers opportunities to analyze the hierarchical relationships of groups of users and gain an understanding of their relative proximity. When we look at the upper levels of level 3, some small communities represented as nodes in block multi-graph show similar connectivity and are grouped together by the SBM model. In level 2, four communities are identified with two climate theory supporter groups, one arson theory supporter group, and one mixed group. In level 1, we identify one small community with arson theory supporters and one big community with the rest of the users. From the hierarchical structure of communities, we notice two different arson theory groups: a stubborn group that is isolated from other groups at all levels and an interactive group that talks to climate theory supporters.

We are particular interested in level 2 where the number of communities and the size of communities are intermediate such that we can analyze each community in detail. Besides the two manual classifications, we identify the origin of the users as being Australian or not. A detailed study of user profiles is listed in Table. 5.2 and shows that the four communities consist of different types of users and can be summarised as international climate group, Australian climate group, Australian mixed group, and international arson group. The special stubborn group consists of famous climate change deniers such as @realDonaldTrump, @RealJamesWoods, and @DonaldJTrumpJr accounts, while climate groups have accounts

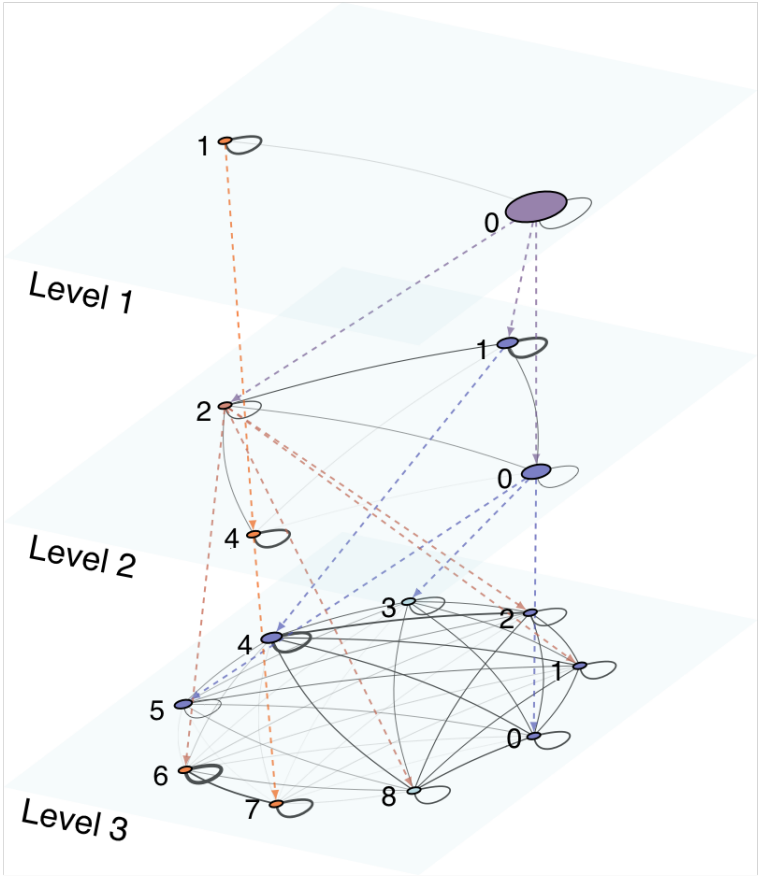
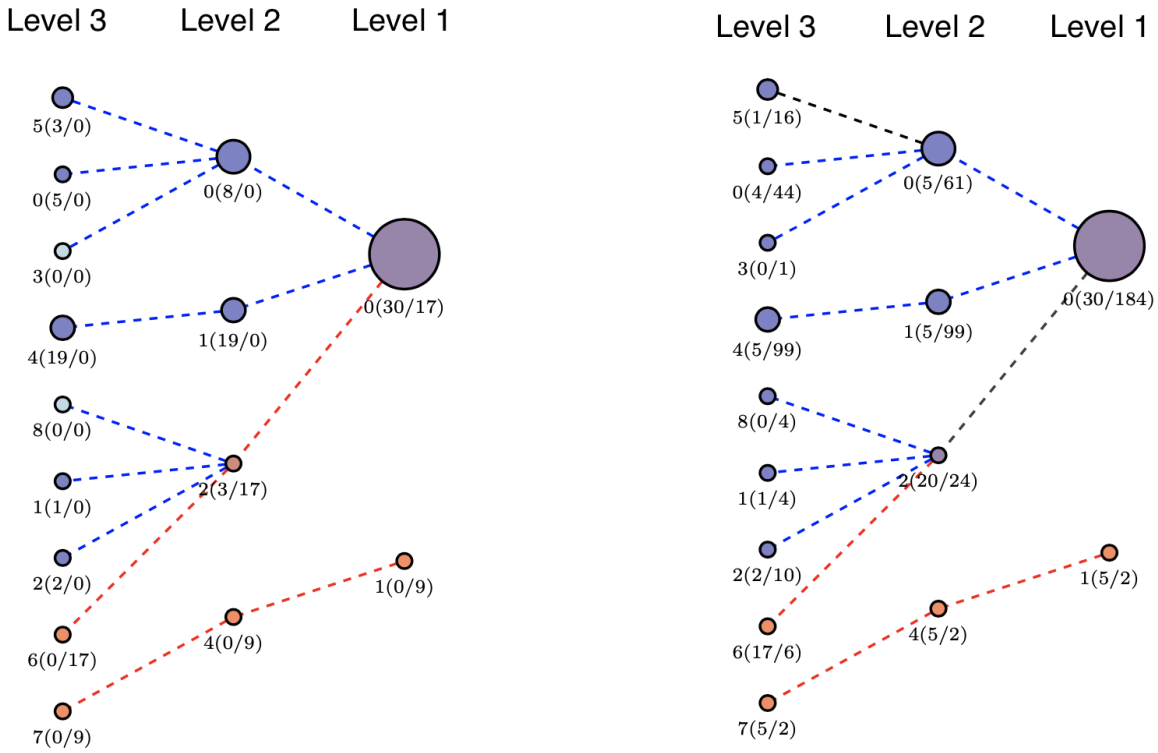


FIGURE 5.3. Top three level of communities from hSBM model based on directed retweet multi-graph. Communities are plotted as nodes in levels where level 1 has two communities, level 2 has four communities, and level 3 has nine communities. Node size represents the corresponding community size, and links within each layer correspond to the density between communities computed by Eq. (3.1). Red and blue dashed links between levels exist if user nodes are assigned in both partitions.

including climate activists and official accounts such as @GretaThunberg, @climatecouncil, @ProfTerryHughes, @MikeCarlton01, @MrKRudd and @GrogsGamut, and mixed group has influential media accounts: @SkyNewsAust and @BOM_au.

After showing the separation of users by hSBM model in level 2, we are also interested in how these communities behave in the bushfire discussion. We answer this question by analyzing the happiness of users from tweets that were split into 4 communities according to partition in level 2 and were collected in a longer timeline from November 2019 to October 2020 with the same selecting criteria listed in Sec. 5.2. In Fig. 5.5, we noticed that all four groups



(A) Arson-Fire (B) Climate-Fire

FIGURE 5.4. Distribution of classified users in the communities shown in Fig. 5.3 based on Arson-Fire classification (left) and Climate-Fire classification (Right) described in Sec. 5.2. We focus on the hierarchical structure of communities between different levels and ignore the retweeting interactions in each level. For each community (represented as a node), classification result is shown in the format: $C_i(N_i^O/N_i^S)$ where N_i^O and N_i^S are the number of opposers and supporters in community C_i, based on the corresponding classifications. The size of node corresponds to the size of the community, and the color of the node corresponds to the majority classification within that community: blue represents communities dominated by users who support the climate view, and red represents those dominated by users who support the arson view. The edge color corresponds to the node color in lower levels.

show a similar trend overall, that is, a decreasing trend in the height stage of bushfire and an increasing trend in the after stage of bushfire. However, they reacted differently in the early stage of bushfires. The international climate group (community 0) and the Australian mixed group (community 2) show increasing happiness, reach a peak at the UN Climate Conference, and drop when approaching the height stage. The international climate group has the highest average happiness among the four groups and reacted positively to climate-supporting events

Community	Name	Nodes	Australian	Influential Users
0	international climate group	1,740	45%	@GretaThunberg @climatecouncil @ProfTerryHughes
1	Australian climate group	1,248	87%	@MikeCarlton01 @MrKRudd @GrogsGamut
2	Australian mixed group	683	80%	@SkyNewsAust @BOM_au @RealMarkLatham
4	international arson group	358	24%	@realDonaldTrump @RealJamesWoods @DonaldJTrumpJr

TABLE 5.2. User features of 4 communities of bushfire retweet network. The columns are the community label (Community), community name (Name) which is manually given by us according to its features, number of nodes in each community (Nodes), fraction of Australian accounts in each community (Australian), and examples of influential users (Influential Users). The partition corresponds to level 2 of hSBM applied to a retweet network.

such as the UN Climate Conference and Earth Day. In contrast, the Australian climate group (community 1) has the lowest average happiness among four communities in three stages of bushfires, showing their concerns about natural disasters. This group shows a strong negative attitude towards bad climate news, such as the protest in Sydney and public suspicion of the climate speech from one candidate of the Liberal National Party (LNP). The most interesting group, the international arson group (community 4) consisting of climate change deniers, shows no increase in climate-supporting events and consistently has relatively low happiness except in one case where climate deniers cheer the climate change speech from Donald Trump. On the same day, Australian users focused on Australian challenges on climate change and thus two Australian groups have a negative mood.

The successful findings in both partitions of hSBM and sentiment analysis confirm the ability of our automated methods in finding distinct groups of users based on their interactions and providing useful insights into the underlying network. We then apply the same approach to reply interaction in the bushfire discussion. The nodes are divided into two communities in

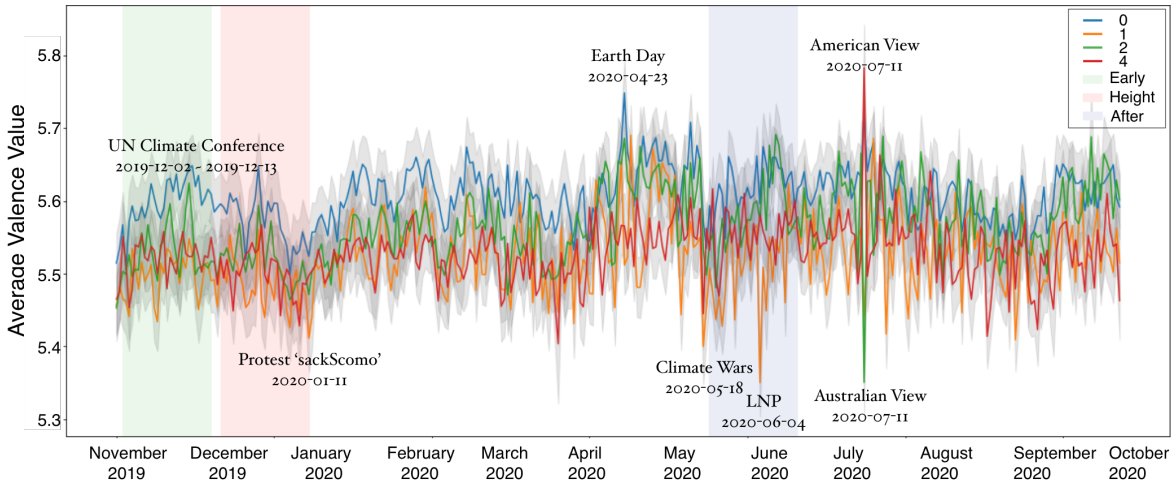
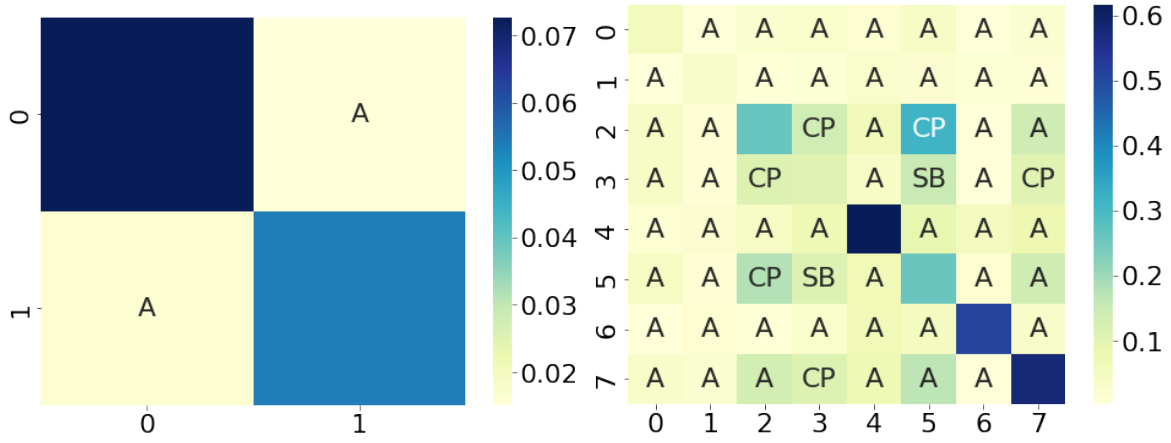


FIGURE 5.5. Sentiment of four communities in bushfire discussion. The average valence value of each community is computed from the tweets containing word "climate" during the time period from November 2019 to October 2020. The standard deviation in grey is computed by the Bootstrapping method same as Sec. 5.3.1.

level 1, eight communities in level 2 as shown in Fig. 5.6, and 21 communities in level 3. The climate theory supporters and arson theory supporters are distributed across different communities even in level 3, indicating that reply activity does not have strong alignment with users' political opinions but show user's participation in political debates, which agrees with previous works on political polarization [Gai+21].

As an important interaction type in online platforms, reply activity provides valuable information about users' involvement in debates. We combine retweet and reply activities and construct a multilayer network with 4,043 nodes and two types of edges. We identify hierarchical communities by the hSBM model, with two communities in level 1, four communities in level 2, and eight communities in level 3 shown in Fig 5.7. Fig. 5.8 shows that climate theory supporters and arson theory supporters are separated into different groups. All classified users based on Arson-Fire and more than 67% classified users based on Climate-Fire share one single opinion within groups. The community structures in Fig 5.7 indicate that users have different connecting preferences between retweet and reply activity and that retweet activity is more assortative(83% assortative fraction in level 2 and 96% in level 3) than reply (67% assortative fraction in level 2 and 64% in level 3). This agrees with the view that retweets



(A) Level 1 (B) Level 2

FIGURE 5.6. Density matrices for the partition obtained by hSBM based on bushfire reply network on level 1 (left) and level 2 (right). The blocks are ranked according to community size $N_r = \{2413, 1541\}$ (A), $N_r = \{1276, 1115, 491, 394, 299, 252, 91, 36\}$ (B). The labels in each entry of the matrix correspond to the structure type τ where "A" stands for "Assortative", "CP" stands for "Core-Periphery", "D" stands for "Disassortative" and "SB" stands for "Source-Basin".

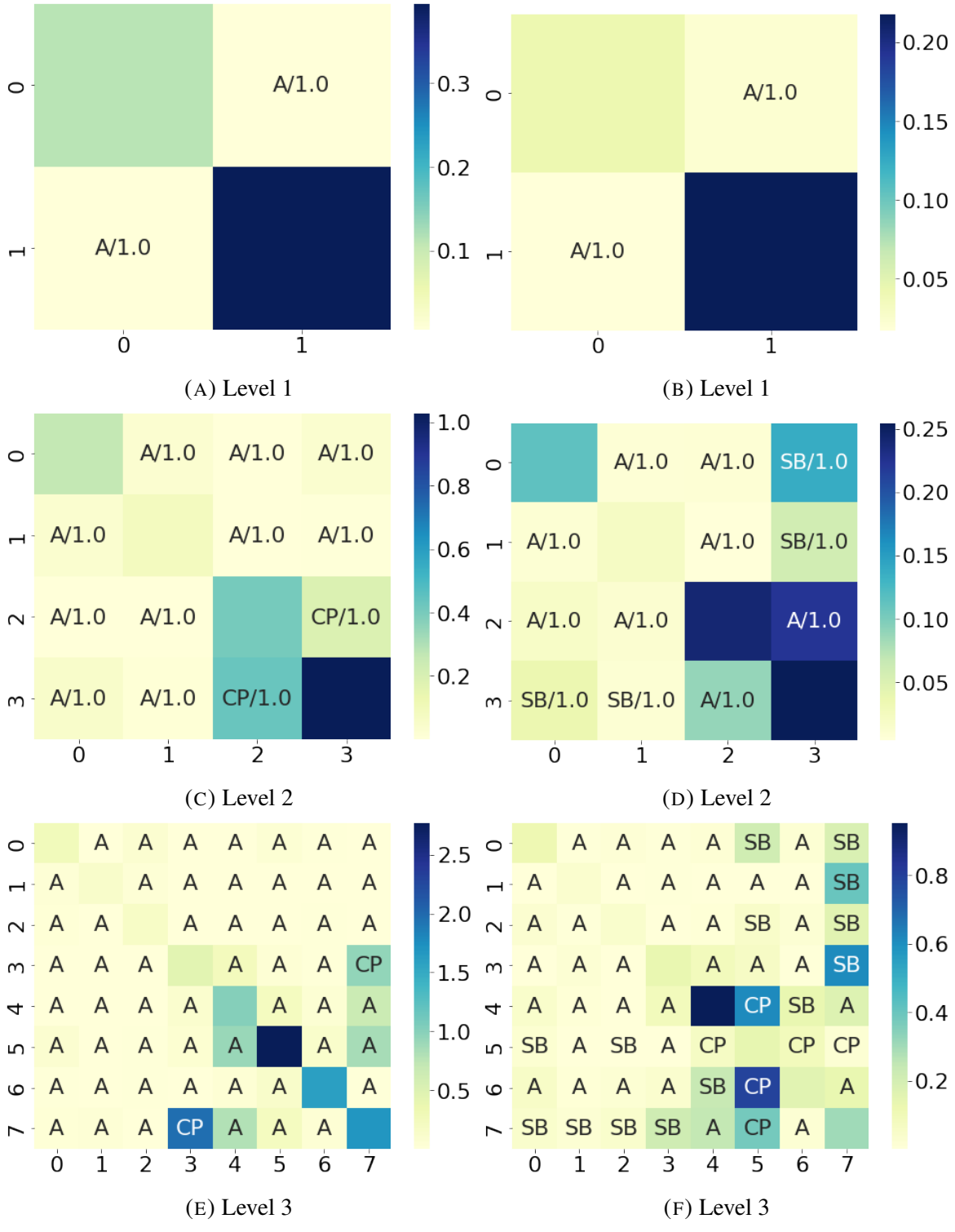
exhibit a highly polarized structure with users receiving information within groups [Con+11a; Cin+21]. Notably, we observe non-assortative structures from level 2 with a core-periphery structure between the arson theory supporter group and a mixed group in retweet network and a source-basin structure between the mixed group and climate theory supporters in reply network. In level 2, the retweeting core group has a high influence in retweet (436 average out-degree) and receives a large quantity of information in reply (591 average in-degree). Compared to information source groups in reply that have mostly climate theory supporters and 33 influential users, the reply basin group has 0 users in the top 1% high influence users. One potential explanation is that the mixed group contains activists of both climate theory supporters and deniers such that the arson group gets news only from this group while climate groups have many information sources and do not rely on this small group. On the other hand, climate change as a popular trend attracts replies from both supporters and deniers. Our findings on multilayer network privide a depper understanding on user's interactions including retweet and reply in political debates, and reveal important interaction structural patterns that have not been reported in previous analysis. It also confirms the potential of our

methodology to identify unusual structures which play important role in the organization and function of underlying networks.

5.4.2 Election

After the successful study on bushfire discussion, we now turn to the analysis of election voting in Germany. Fig. 5.9 shows the communities obtained from hSBM on election networks are dominant by assortative structure and other structure types only appear in level 3 in both retweet and reply networks. A detailed study on retweet communities in level 3 indicates that a group with left-leaning politicians and activists such as Grüne, ParentsForFuture, Linke, SPD (community 1) acts as core and is retweeted a lot by a public periphery (community 0) in which only two users among 11,375 users from Antifa and Linke are identified. In contrast, a source-basin structure appears within right-wing parties. The source (community 3) includes many politicians with eight from Afd and others from Freie Wähler and Blaue Partei, while the basin group (community 2) has many political activists from Pegida and Identitäre, and two politicians from Der III. Weg and Freie Wähler. This finding demonstrates that the left-leaning and right-leaning parties have different interaction patterns. It also reveals important insights about the different nature between core-periphery and source-basin that core-periphery is found between densely connected influential users and public, while source-basin is obtained between active users and relative loosely connected influential users. The reply network, however, could not find groups with different political opinions which confirms the statement that reply activity is visible to all users and different opinions are possible to confront each other while retweeting mainly happens from followers [Gai+21].

We continue our analysis using multilayer network, that includes retweet and reply interactions. The communities by hSBM in Fig. 5.10 confirm the difference of interaction patterns between the user's retweet and reply activity. While most pairwise interactions between communities are assortative in all levels, core-periphery structure can be obtained in level 2 in both retweet and reply and source-basin structure starts appearing in a more fine-grained level. However, the political parties of users are a bit mixed into groups in level 2 which might be due to the fact that reply activity across different groups perturbs the identification of political groups.

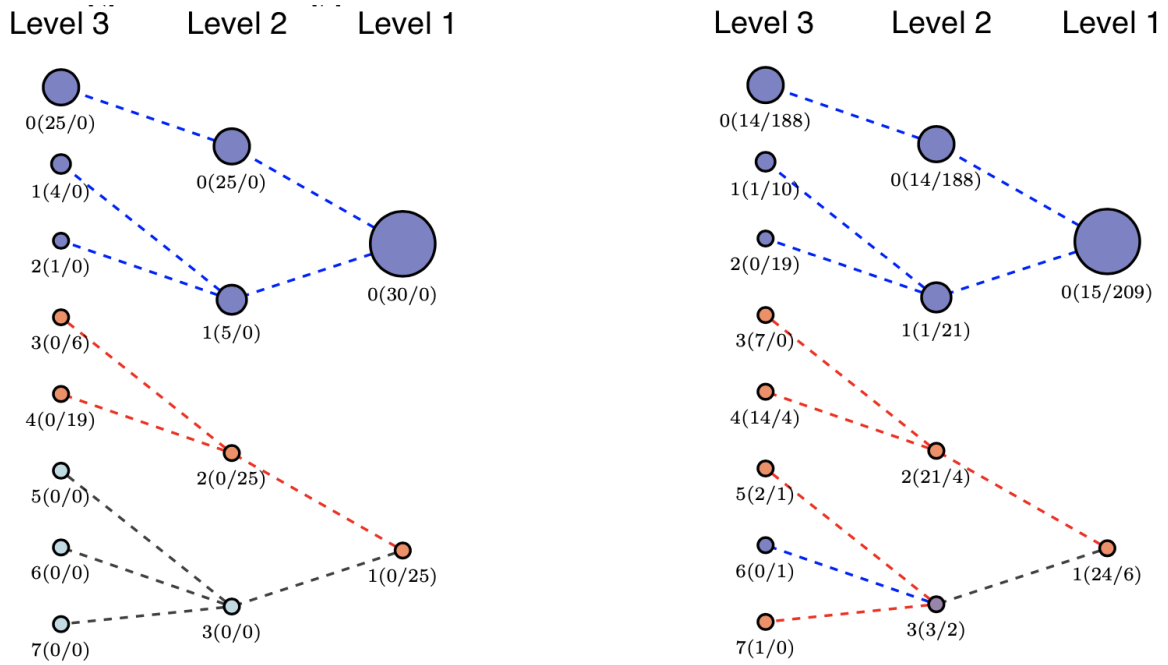


(A) Level 1 (B) Level 1

(C) Level 2 (D) Level 2

(E) Level 3 (F) Level 3

FIGURE 5.7. Density matrices for the top 3 levels of communities obtained by hSBM based on bushfire multilayer network according to retweet interaction (Left) and reply interaction (Right). The blocks are ranked according to community size $N_r = \{3421, 622\}$ (A,B), $N_r = \{1870, 1551, 522, 100\}$ (C,D), $N_r = \{1870, 1001, 550, 315, 207, 49, 36, 15\}$ (E,F). The labels in each entry of the matrix correspond to the structure type τ and the robustness score obtained from Eqn. (A.1).



(A) (B)

FIGURE 5.8. Distribution of classified users in the communities shown in Fig. 5.7 based on Arson-Fire classification (left) and Climate-Fire classification (Right) described in Sec. 5.2. We focus on the hierarchical structure of communities between different levels and ignore the retweeting interactions in each level. For each community (represented as a node), classification result is shown in the format: $C_i(N_i^O/N_i^S)$ where N_i^O and N_i^S are the number of opposers and supporters in community C_i, based on the corresponding classifications. The size of node corresponds to the size of the community, and the color of the node corresponds to the majority classification within that community: blue represents communities dominated by users who support the climate view, and red represents those dominated by users who support the arson view. The edge color corresponds to the node color in lower levels.

This finding indicates that the user's whole activity in online social media is not strongly polarized at a loose-grained level and that a finer partition should be considered.

5.4.3 New Year Eve(NYE)

The last dataset we used in our project is New Year Eve attack in Germany. We apply hSBM on three networks of this dataset: retweet, reply and multilayer with retweet and reply. The community structures obtained based on a single interaction type (see Appendix Fig. C.3)

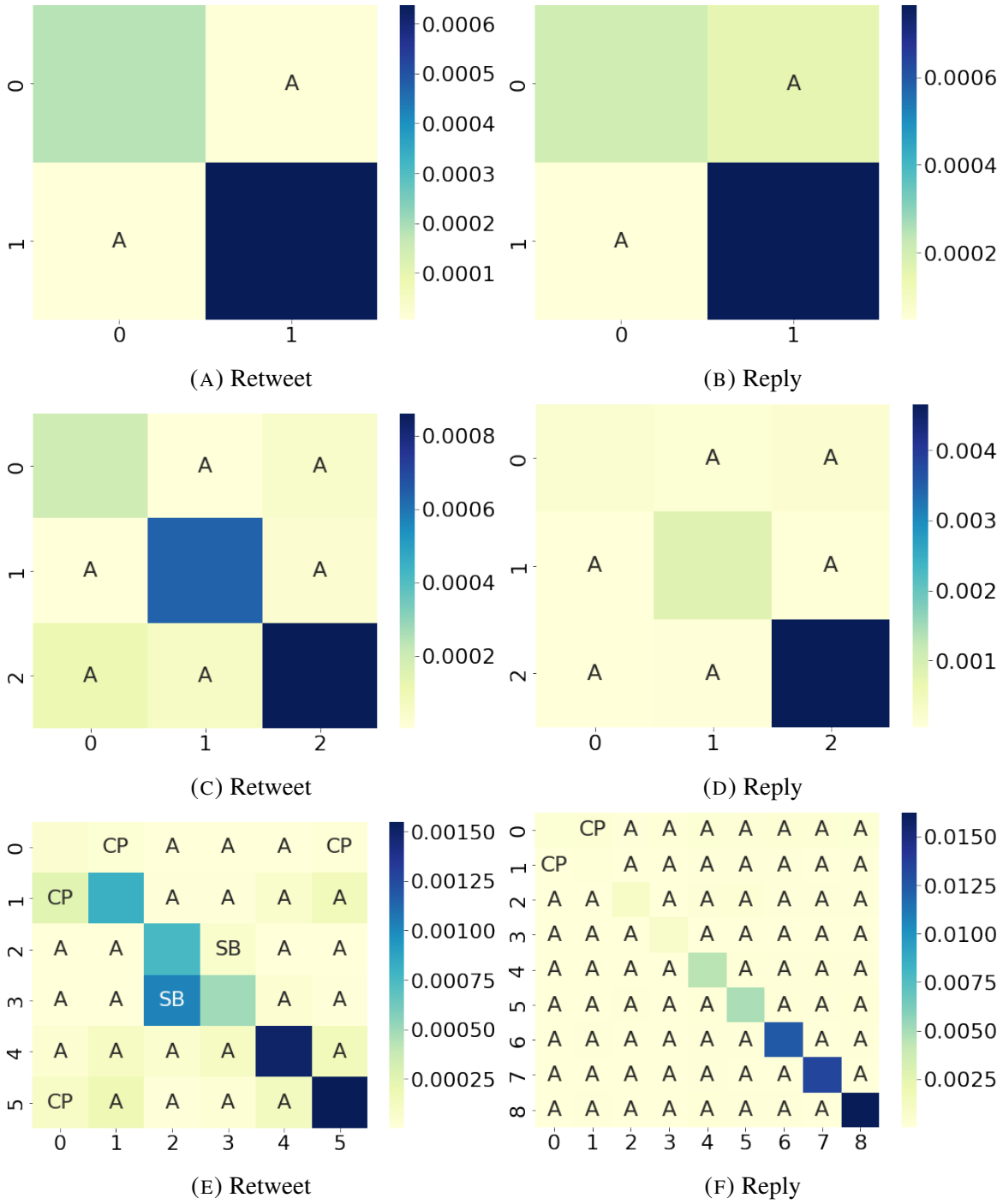
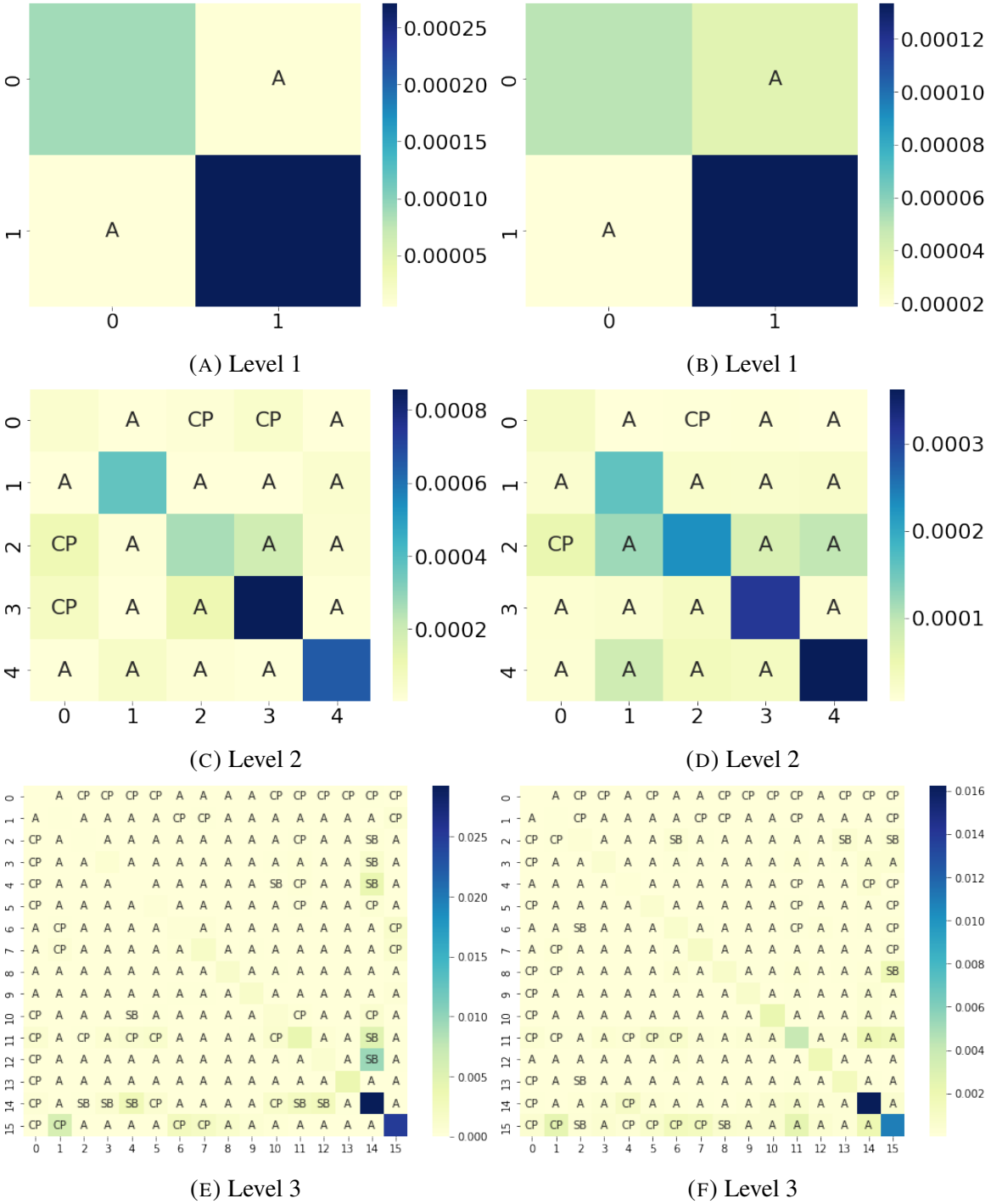


FIGURE 5.9. Density matrices for the top 3 levels of communities obtained by hSBM based on election retweet network (Left) and reply network (Right). The blocks are ranked according to community size $N_r = \{21743, 7822\}$ (A), $N_r = \{18150, 7822, 3593\}$ (C), $N_r = \{11375, 6775, 5398, 2424, 2094, 1490\}$ (E), $N_r = \{15952, 4286\}$ (B), $N_r = \{15952, 3351, 935\}$ (D), $N_r = \{6134, 5961, 3857, 1661, 924, 766, 365, 301, 269\}$ (F). The labels in each entry of the matrix correspond to the structure type τ.



(A) Level 1 (B) Level 1

(C) Level 2 (D) Level 2

(E) Level 3 (F) Level 3

FIGURE 5.10. Density matrices for the top 3 levels of communities obtained by hSBM based on election multilayer network according to retweet interaction (Left) and reply interaction (Right). The blocks are ranked according to community size $N_r = \{28617, 12219\}$ (A,B), $N_r = \{16458, 9460, 8046, 4113, 2759\}$ (C,D), $N_r = \{12966, 5627, 4152, 3326, 3215, 2556, 1969, 1688, 1200, 1038, 936, 568, 566, 521, 332, 176\}$ (E,F). The labels in each entry of the matrix correspond to the structure type τ.

shows that communities have mostly assortative interactions with others and a small fraction of non-assortative interactions is obtained in retweet (33% in level 2 and 29% in level 3) and reply (7% in level 2 and 23% in level 3). The communities of multilayer network in Fig. C.4 also show the appearance of core-periphery structures with a larger fraction in retweet than reply. However, in this dataset, the communities found in all three networks – retweet, reply and multilayer – do not show a clear separation of political parties. It suggests a deeper understanding of this dataset is needed.

5.5 Discussion

The large amount of information in social media is beneficial for the analysis of various problems, while the complexity of social media data brings challenges. In this chapter, we took a quantitative analysis of social media data including public happiness measurement from textual data and group-level relationships analysis from the social interactions of users based on network topology. The sentiment analysis on the bushfire dataset provides an understanding of bushfire discussion, revealing that the public has the lowest happiness in the height stage of bushfire, and climate change supporters have increasing happiness when climate-supporting events occur while climate change deniers do not change their mood.

The analysis of social structures shows that the partitions found in retweet networks are strongly polarized in bushfire and election events while reply networks do not show a clear pattern in political view. This is in line with the interpretation that retweets align with political opinions and replies illustrate user's participation. Notably, core-periphery or source-basin structures are identified in almost all networks, confirming the widespread appearance of non-assortative structures in social networks. In particular, we identified a group of climate activists in bushfire discussion who influence deniers of climate change theory through retweets and frequently engage with climate theory supporters through replies. The data of election voting in Germany reveals distinct interaction patterns within the left-wing and right-wing parties. The left-leaning party exhibits a core-periphery structure with the public, while the right-leaning party displays an internal source-basin configuration with more politicians in

source and more political activists in basin. The detailed study on source-basin reveals a new type of information flow from loosely connected influential users to actively interactive users in social networks, which is different from classical core-periphery structures. The evidence of the appearance of unusual community structures in networks that have not been reported in previous analyses confirms the potential of our methodology developed in Chapter 3 in the analysis of complex networks.

Multilayer networks have potential of considering multiple interaction types which are common in social media, and give a richer picture of user's interaction. In theory, multilayer approach can provide information in more dimensions and find groups more successful than a single layer. But in practice, we find that multilayer networks can find groups with political preferences but users in different parties are a bit mixed. It might be the reason that replies are not restricted to their own parties so that when we detect community based on retweets and replies, the partition is not aligned with political affiliation perfectly as retweets. On the other hand, the advantage of a multilayer is that users are partitioned considering multifaceted features and a comprehensive analysis is possible. Multilayer feature also provides the same standard for the comparison between multiple types of information.

CHAPTER 6

Conclusion

6.1 Summary of the main results

The information era brings us an enormous quantity of data that provides valuable information for the analysis of social, biological, transportation, and financial systems. The network approach has become a convenient and easy-to-fit tool to gain an understanding of relationships and interactions between elements of such complex systems. While mesoscale structures in complex networks have been successfully explored using community-detection methods for more than two decades, the majority of studies assume communities to have either assortative or core-periphery connectivity patterns. In this project, we show that these two structures are not all possibilities of connectivity patterns, and other structures are mathematically possible and statistically typical in complex networks, as long as we use algorithms that are agnostic about the type of structural pattern to detect communities. We introduce a methodology to identify community structure types in undirected or directed multi-graphs. In particular, through pairwise comparison on the connectivity strength between communities, we classify four community structures in directed graphs: assortative(A), core-periphery(CP), disassortative(D), and especially a new type of community structure, source-basin(SB), in which information(influence) flows from sparsely connected source group to a strongly connected basin group of information receivers. The newly introduced source-basin indicates another type of organization of networks that has not been discovered in previous analysis and that universally appear in all categories of empirical networks extracted from the Netzschleuder repository.

Despite the success of identifying all community structures within empirical networks, the mechanism leading to the formation of these pair-wise structures is often unclear. To address this problem, we proposed a simple generative model to mimic the evolution process of networks with two communities. With the assumption that nodes have agency in determining their incoming edges (who they are influenced by) but not their outgoing edges (who they influence), the model parameter regimes show the necessary conditions for the appearance of the four structure types and reveal possible mechanisms of the formation of these structures. In particular, we find that core-periphery structure cannot be obtained if two groups have the same preference on link assortativity, and source-basin structure only appears when two groups have different degrees. Moreover, under in-degree variation condition, source-basin appears when the link assortativity of both groups is close to a neutral state, and increasing (decreasing) link assortativity in both groups leads to assortative (disassortative) structures, and asymmetric assortativity results in core-periphery. In general, variations in group sizes and average degree facilitate the appearance of source-basin and tighten the conditions of other structures.

Our generative model shows a close relationship between source-basin structure and the difference in the in-degree of the two groups. The group with a higher in-degree is identified as basin, who is influenced more by the source group with a lower in-degree possibly because they are more restrictive about information sources. Moreover, source-basin is the only arrangement when users only change their influence source numbers during the evolution process, showing that when the only difference between two groups is in-degree, only source-basin relationship can be observed. The source is the group that receives less influence and the basin is the group that is influenced more. In the case of two groups discussed in Sec. 4, we derived that a degree variability between the groups is a necessary condition for the appearance of SB and CP structures (independent of the generative model). The close relation between core-periphery and degree difference was shown in previous analysis on undirected networks [BL16; ZMN15]. No core-periphery structure can be found in the two groups if we use a degree correction method [KM18]. In the general case with more than two groups, no such constraints apply and all four community types are possible when applying a pair-wise classification (as done in [LAA23]).We further clarify the connection between degree and structure type in the directed network by using an alternative definition of density matrix

that takes degree into account. We find that neither source-basin nor core-periphery can be obtained with the new density matrix, while other community structure types can appear.

Apart from the theoretical developments in classification and network evolution model of community structure, we further investigate the significant role that community structures play in the organization of networks through a quantitative analysis of social media data, taking the advantage of interpretable user interactions in social media platform X that generates vast quantities of information. In political debate events in Australia and Germany, we find that retweet networks are very polarized, with distinct groups of users showing different political preferences. The community structures found in bushfire and election datasets reveal important mesoscopic pattern of user' interaction. In particular, the bushfire has a mixed group with climate activists who influence climate deniers through retweet and frequently engage with climate supporters through reply. The election voting patterns in Germany reveal distinct interaction patterns within the left-wing and right-wing parties.

The evidence of less-studied community structure types in almost all networks reveals a new picture of organization in online networks. The source group acts as a source of information, containing users that are more central and do not interact significantly with each other. The basin group contains active users who reference both source nodes and each other. This is different from the well-studied core-periphery case where influential users actively interact and are well-connected with each other, while peripheral users do not communicate with one another and receive most of their influence from the core group. Despite the successful findings in social networks, the methodology we proposed can be applied to any directed multi-graphs and we expect similar results can be found in studies of other areas. The fact that such community patterns have not been previously reported in these cases, despite being extensively studied networks and construction approaches, confirms both the potential of our methodology and the limitations of community-detection methods that fix a prior specific structures. We can thus expect that studies with networks in other areas will be similarly successful in identifying less-studied structures and providing new insights not only on the role of specific nodes and communities but also on the organization and function of the network as a whole.

6.2 Future outlook

Beyond the generalization of mesoscale structure type in directed networks, future work should consider how to extend the applicability and interpretability of our methods and findings. For instance, generalizations that overcome simplifying assumptions of our methods could consider more general types networks (e.g., multi-layered) and definitions of community type (e.g., beyond pairwise classification).

The generative model also performs on a simple setting, suggesting different extensions of our model to consider more general cases and more realistic network-evolution processes. For instance, one could consider local searches in the network when re-wiring and also processes in which the group affiliation of nodes changes over time (depending on the neighbors). An important extension would be to consider more than two groups, as this would lift many of the restrictions on the appearance of SB and, therefore, potentially show a richer picture for pairwise SB relationships. With more than two communities, a pairwise SB relationship can occur because of the excess degree that the two communities acquire due to different interactions with other communities. A direct application of these models to data would require not only a combination of these different extensions but also longitudinal data to distinguish between the rewiring preferences of different nodes.

The analysis on social media data suggests many improvements. For example, the mixture of political parties found with multiple sources of information suggests that a finer community level should be considered, one in which users are divided into relatively small communities with specific features in both retweets and replies. The small mixed group in bushfire discussion, which has interesting interaction patterns in retweets and replies, suggests a detailed study on the users within this group. The pattern found in NYE dataset also suggests a deeper analysis on online user's involvement in this violence event. For sentiment analysis, and interesting further exploration would be detailed study on the words used in different communities and "word shift" graphs by Gallagher et al (2021) could be considered.

Bibliography

[AB02] Réka Albert and Albert-László Barabási. 'Statistical mechanics of complex networks'. In: *Reviews of modern physics* 74.1 (2002), p. 47.

[AG05] Lada A Adamic and Natalie Glance. 'The political blogosphere and the 2004 US election: divided they blog'. In: *Proceedings of the 3rd international workshop on Link discovery*. 2005, pp. 36–43.

[Ayn20] Thomas Aynaud. *python-louvain x.y: Louvain algorithm for community detection*. 2020. URL: https://github.com/taynaud/python-louvain.

[Bar13] Albert-László Barabási. 'Network science'. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 371.1987 (2013), p. 20120375.

[BBB18] Richard F Betzel, Maxwell A Bertolero and Danielle S Bassett. 'Non-assortative community structure in resting and task-evoked functional brain networks'. In: *Biorxiv* (2018), p. 355016.

[BE00] Stephen P Borgatti and Martin G Everett. 'Models of core/periphery structures'. In: *Social networks* 21.4 (2000), pp. 375–395.

[Bed+22] Monika Bednarek et al. 'Winning the discursive struggle? The impact of a significant environmental crisis event on dominant climate discourses on Twitter'. In: *Discourse, Context & Media* 45 (2022), p. 100564.

[BJC17] Matthew C Benigni, Kenneth Joseph and Kathleen M Carley. 'Online extremism and the communities that sustain it: Detecting the ISIS supporting community on Twitter'. In: *PloS one* 12.12 (2017), e0181405.

[BL16] Paolo Barucca and Fabrizio Lillo. 'Disentangling bipartite and core-periphery structure in financial networks'. In: *Chaos, Solitons & Fractals* 88 (2016), pp. 244–253.

[Blo+08] Vincent D Blondel et al. 'Fast unfolding of communities in large networks'. In: *Journal of statistical mechanics: theory and experiment* 2008.10 (2008), P10008.

[Bol91] Marianna Bolla. *Relations between spectral and classification properties of multigraphs*. DIMACS, Center for Discrete Mathematics and Theoretical Computer Science, 1991.

[BU89] Daniel Baird and Robert E Ulanowicz. 'The seasonal dynamics of the Chesapeake Bay ecosystem'. In: *Ecological monographs* 59.4 (1989), pp. 329–364.

[Car+13] Alessio Cardillo et al. 'Emergence of network features from multiplexity'. In: *Scientific reports* 3.1 (2013), p. 1344.

[CH90] Kenneth Church and Patrick Hanks. 'Word association norms, mutual information, and lexicography'. In: *Computational linguistics* 16.1 (1990), pp. 22–29.

[Cin+21] Matteo Cinelli et al. 'The echo chamber effect on social media'. In: *Proceedings of the National Academy of Sciences* 118.9 (2021), e2023301118.

[CLX16] Shaosheng Cao, Wei Lu and Qiongkai Xu. 'Deep neural networks for learning graph representations'. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 30. 1. 2016.

[CM15] Darko Cherepnalkoski and Igor Mozetic. 'A retweet network analysis of the European Parliament'. In: *2015 11TH International Conference On Signal-Image Technology & Internet-Based Systems (SITIS)*. IEEE. 2015, pp. 350–357.

[CMN08] Aaron Clauset, Cristopher Moore and Mark EJ Newman. 'Hierarchical structure and the prediction of missing links in networks'. In: *Nature* 453.7191 (2008), pp. 98–101.

[CNM04] Aaron Clauset, Mark EJ Newman and Cristopher Moore. 'Finding community structure in very large networks'. In: *Physical review E* 70.6 (2004), p. 066111.

[Cod+15] Emily M Cody et al. 'Climate change sentiment on Twitter: An unsolicited public opinion poll'. In: *PloS one* 10.8 (2015), e0136092.

[Con+11a] Michael Conover et al. ‘Political polarization on twitter’. In: *Proceedings of the international aaai conference on web and social media*. Vol. 5. 1. 2011, pp. 89–96.

[Con+11b] Michael D Conover et al. ‘Predicting the political alignment of twitter users’. In: *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*. IEEE. 2011, pp. 192–199.

[Con+12] Rosaria Conte et al. ‘Manifesto of computational social science’. In: *The European Physical Journal Special Topics* 214 (2012), pp. 325–346.

[Dan20] Martin Rosvall Daniel Edler Anton Eriksson. *Infomap network clustering algorithm*. 2020. URL: https://pypi.org/project/infomap/.

[DCT19] Lorenzo Dall’Amico, Romain Couillet and Nicolas Tremblay. ‘Revisiting the bethe-hessian: improved community detection in sparse heterogeneous graphs’. In: *Advances in neural information processing systems* 32 (2019).

[Ell+20] Andrew Elliott et al. ‘Core–periphery structure in directed networks’. In: *Proceedings of the Royal Society A* 476.2241 (2020), p. 20190783.

[Esp+22] Lisette Espín-Noboa et al. ‘Inequality and Inequity in Network-Based Ranking and Recommendation Algorithms’. In: *Scientific Reports* 12.1 (Feb. 2022), p. 2012. ISSN: 2045-2322.

[FN22] Santo Fortunato and Mark EJ Newman. ‘20 years of network community detection’. In: *Nature Physics* 18.8 (2022), pp. 848–850.

[For10] Santo Fortunato. ‘Community detection in graphs’. In: *Physics reports* 486.3-5 (2010), pp. 75–174.

[FWW17] Yanghe Feng, Qi Wang and Tengjiao Wang. ‘Drug target protein-protein interaction networks: a systematic perspective’. In: *BioMed research international* 2017.1 (2017), p. 1289259.

[Gai+21] Felix Gaisbauer et al. ‘Ideological differences in engagement in public debate on Twitter’. In: *PLOS ONE* 16.3 (25th Mar. 2021). Publisher: Public Library of Science, e0249241.

[GDB06] Thilo Gross, Carlos J Dommar D'Lima and Bernd Blasius. 'Epidemic dynamics on an adaptive network'. In: *Physical review letters* 96.20 (2006), p. 208701.

[Gle+14] James P Gleeson et al. 'A simple generative model of collective online behavior'. In: *Proceedings of the National Academy of Sciences* 111.29 (2014), pp. 10411–10415.

[GN02] M. Girvan and M. E. J. Newman. 'Community Structure in Social and Biological Networks'. In: *Proceedings of the National Academy of Sciences* 99.12 (2002), p. 7821.

[GN05] Roger Guimera and Luıés A Nunes Amaral. 'Functional cartography of complex metabolic networks'. In: *nature* 433.7028 (2005), pp. 895–900.

[GS09] Thilo Gross and Hiroki Sayama. *Adaptive Networks: Theory, Models and Applications*. en. Springer Science & Business Media, Aug. 2009. ISBN: 978-3-642-01284-6.

[GSA07] Roger Guimera, Marta Sales-Pardo and Luis AN Amaral. 'Classes of complex networks defined by role-to-role connectivity profiles'. In: *Nature physics* 3.1 (2007), pp. 63–69.

[GYW21] Ryan J Gallagher, Jean-Gabriel Young and Brooke Foucault Welles. 'A clarified typology of core-periphery structure in networks'. In: *Science advances* 7.12 (2021), eabc9800.

[HLL83] Paul W Holland, Kathryn Blackmond Laskey and Samuel Leinhardt. 'Stochastic blockmodels: First steps'. In: *Social networks* 5.2 (1983), pp. 109–137.

[HPZ11] Adam Douglas Henry, Paweł Prałat and Cun-Quan Zhang. 'Emergence of Segregation in Evolving Social Networks'. In: *Proceedings of the National Academy of Sciences* 108.21 (May 2011), pp. 8605–8610. ISSN: 0027-8424, 1091-6490.

[KE02] Konstantin Klemm and Víctor M. Eguíluz. 'Growing Scale-Free Networks with Small-World Behavior'. In: *Physical Review E* 65.5 (May 2002), p. 057102.

[Kiv+14] Mikko Kivelä et al. 'Multilayer networks'. In: *Journal of complex networks* 2.3 (2014), pp. 203–271.

[KM17] Sadamori Kojaku and Naoki Masuda. 'Finding multiple core-periphery pairs in networks'. In: *Physical Review E* 96.5 (2017), p. 052313.

[KM18] Sadamori Kojaku and Naoki Masuda. 'Core-periphery structure requires something else in the network'. In: *New Journal of physics* 20.4 (2018), p. 043012.

[KN11] Brian Karrer and Mark EJ Newman. 'Stochastic blockmodels and community structure in networks'. In: *Physical review E* 83.1 (2011), p. 016107.

[LAA23] Cathy Xuanchi Liu, Tristram J Alexander and Eduardo G Altmann. 'Nonassortative relationships between groups of nodes are typical in complex networks'. In: *PNAS nexus* 2.11 (2023), pgad364.

[Laz+09] David Lazer et al. 'Computational social science'. In: *Science* 323.5915 (2009), pp. 721–723.

[Lew+10] Anna CF Lewis et al. 'The function of communities in protein interaction networks at multiple scales'. In: *BMC systems biology* 4 (2010), pp. 1–14.

[LF09] Andrea Lancichinetti and Santo Fortunato. 'Community detection algorithms: a comparative analysis'. In: *Physical review E* 80.5 (2009), p. 056117.

[LG14] Omer Levy and Yoav Goldberg. 'Neural word embedding as implicit matrix factorization'. In: *Advances in neural information processing systems* 27 (2014).

[Liu+20] Fanzhen Liu et al. 'Deep learning for community detection: progress, challenges and opportunities'. In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*. Vol. IJCAI-20. 2020, p. 4982.

[Liu23] Cathy Liu. *Repository of codes and data*. Version 1.0.0. Jan. 2023. URL: https://github.%20sydney.edu.au/xliu0689/CommunityStructureType.

[Liu24] Cathy Liu. *Repository with codes.* Available on GitHub https://github.com/communityType/model and (an archived version) on Zenodo [accessed 2024 Sep] https://doi.org/10.5281/zenodo.13846866. 2024.

[LK77] J. Richard Landis and Gary G. Koch. 'The Measurement of Observer Agreement for Categorical Data'. In: *Biometrics* 1.Oct (1977), pp. 159–174.

[Mac03] David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[McH12] Mary L McHugh. 'Interrater reliability: the kappa statistic'. In: *Biochemia medica* 22.3 (2012), pp. 276–282.

[Mit+13] Lewis Mitchell et al. 'The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place'. In: *PloS one* 8.5 (2013), e64417.

[Moo+11] Cristopher Moore et al. 'Active learning for node classification in assortative and disassortative networks'. In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2011, pp. 841–849.

[MPA24] Diogo Melo, Luisa F Pallares and Julien F Ayroles. 'Reassessing the modularity of gene co-expression networks using the Stochastic Block Model'. In: *PLoS computational biology* 20.7 (2024), e1012300.

[Muc+10] Peter J Mucha et al. 'Community structure in time-dependent, multiscale, and multiplex networks'. In: *science* 328.5980 (2010), pp. 876–878.

[MWW20] Elena Milani, Emma Weitkamp and Peter Webb. 'The visual vaccine debate on Twitter: A social network analysis'. In: *Media and Communication* 8.2 (2020), pp. 364–375.

[New03] M. E. J. Newman. 'The Structure and Function of Complex Networks'. In: *SIAM Review* 45.2 (2003), pp. 167–256.

[New06] Mark EJ Newman. 'Modularity and community structure in networks'. In: *Proceedings of the national academy of sciences* 103.23 (2006), pp. 8577–8582.

[New18] Mark Newman. *Networks*. Oxford university press, 2018.

[NL07] Mark EJ Newman and Elizabeth A Leicht. 'Mixture models and exploratory analysis in networks'. In: *Proceedings of the National Academy of Sciences* 104.23 (2007), pp. 9564–9569.

[OM+09] Jukka-Pekka Onnela, Peter J Mucha et al. 'Communities in networks'. In: (2009).

[Pee15] Leto Peel. 'Active discovery of network roles for predicting the classes of network nodes'. In: *Journal of Complex Networks* 3.3 (2015), pp. 431–449.

[Pee17] Leto Peel. 'Graph-based semi-supervised learning for relational networks'. In: *Proceedings of the 2017 SIAM international conference on data mining*. SIAM. 2017, pp. 435–443.

[Pei] Tiago P. Peixoto. *The Netzschleuder network catalogue and repository*. URL: https://networks.skewed.de/ (visited on 2020).

[Pei13] Tiago P Peixoto. 'Parsimonious module inference in large networks'. In: *Physical review letters* 110.14 (2013), p. 148701.

[Pei14a] Tiago P Peixoto. 'Efficient Monte Carlo and greedy heuristic for the inference of stochastic block models'. In: *Physical Review E* 89.1 (2014), p. 012804.

[Pei14b] Tiago P Peixoto. 'Hierarchical block structures and high-resolution model selection in large networks'. In: *Physical Review X* 4.1 (2014), p. 011047.

[Pei17] Tiago P Peixoto. 'Nonparametric Bayesian inference of the microcanonical stochastic block model'. In: *Physical Review E* 95.1 (2017), p. 012317.

[Pei19] Tiago P Peixoto. 'Bayesian stochastic blockmodeling'. In: *Advances in network clustering and blockmodeling* (2019), pp. 289–332.

[Pei20] Tiago P Peixoto. *The graph-tool python library*. 2020. URL: https://graph-tool.skewed.de.

[Pei21] Tiago P Peixoto. 'Revealing consensus and dissensus between network partitions'. In: *Physical Review X* 11.2 (2021), p. 021003.

[Pow+11] Jonathan D Power et al. 'Functional network organization of the human brain'. In: *Neuron* 72.4 (2011), pp. 665–678.

[RB08] Martin Rosvall and Carl T Bergstrom. 'Maps of random walks on complex networks reveal community structure'. In: *Proceedings of the national academy of sciences* 105.4 (2008), pp. 1118–1123.

[RD03] Fritz Jules Roethlisberger and William J Dickson. *Management and the Worker*. Vol. 5. Psychology press, 2003.

[Red+11] Veronica Red et al. 'Comparing community structure to characteristics in online collegiate social networks'. In: *SIAM review* 53.3 (2011), pp. 526–543.

[Red98] Sidney Redner. ‘How popular is your paper? An empirical study of the citation distribution’. In: *The European Physical Journal B-Condensed Matter and Complex Systems* 4.2 (1998), pp. 131–134.

[Rom+17] Puck Rombach et al. ‘Core-periphery structure in networks (revisited)’. In: *SIAM review* 59.3 (2017), pp. 619–646.

[Ros+19] Martin Rosvall et al. ‘Different approaches to community detection’. In: *Advances in network clustering and blockmodeling* (2019), pp. 105–119.

[Roz21] Benedek Rozemberczki. *awesome-community-detection*. https://github.com/benedekrozemberczki/awesome-community-detection?tab=coc-ov-file. 2021.

[Say15] Hiroki Sayama. *Introduction to the modeling and analysis of complex systems*. Open SUNY Textbooks, 2015.

[Sha48] C. E. Shannon. ‘A mathematical theory of communication’. In: *The Bell System Technical Journal* 27.3 (July 1948). Conference Name: The Bell System Technical Journal, pp. 379–423.

[SKZ14] Alaa Saade, Florent Krzakala and Lenka Zdeborová. ‘Spectral clustering of graphs with the bethe hessian’. In: *Advances in Neural Information Processing Systems* 27 (2014).

[Sma+08] Michael Small et al. ‘Scale-Free Networks Which Are Highly Assortative but Not Small World’. In: *Physical Review E* 77.6 (June 2008), p. 066112.

[Su+22] Xing Su et al. ‘A comprehensive survey on community detection with deep learning’. In: *IEEE Transactions on Neural Networks and Learning Systems* (2022).

[The+10] Mike Thelwall et al. ‘Sentiment strength detection in short informal text’. In: *Journal of the American Society for Information Science and Technology* 62 (Dec. 2010), pp. 419–419.

[TL11] Lei Tang and Huan Liu. ‘Leveraging social media networks for classification’. In: *Data mining and knowledge discovery* 23 (2011), pp. 447–478.

[Ure+23] Javier Ureña-Carrión et al. ‘Assortative and preferential attachment lead to core-periphery networks’. In: *Physical Review Research* 5.4 (2023), p. 043287.

[Van+14] Iman Van Lelyveld et al. 'Finding the core: Network structure in interbank markets'. In: *Journal of Banking & Finance* 49 (2014), pp. 27–40.

[Ver+16] Trivik Verma et al. 'Emergence of core–peripheries in networks'. In: *Nature communications* 7.1 (2016), p. 10441.

[Von07] Ulrike Von Luxburg. 'A tutorial on spectral clustering'. In: *Statistics and computing* 17 (2007), pp. 395–416.

[Wes+01] Douglas Brent West et al. *Introduction to graph theory*. Vol. 2. Prentice hall Upper Saddle River, 2001.

[WKB13] Amy Beth Warriner, Victor Kuperman and Marc Brysbaert. 'Norms of valence, arousal, and dominance for 13,915 English lemmas'. In: *Behavior Research Methods* 45 (Feb. 2013), pp. 1191–1207.

[WS98] Duncan J Watts and Steven H Strogatz. 'Collective dynamics of 'small-world'networks'. In: *nature* 393.6684 (1998), pp. 440–442.

[WT07] Ken Wakita and Toshiyuki Tsurumi. 'Finding community structure in mega-scale social networks: [extended abstract]'. In: *Proceedings of the 16th international conference on World Wide Web*. WWW '07. New York, NY, USA: Association for Computing Machinery, 8th May 2007, pp. 1275–1276.

[XS04] R. Xulvi-Brunet and I. M. Sokolov. 'Reshuffling Scale-Free Networks: From Random to Assortative'. In: *Physical Review E* 70.6 (Dec. 2004), p. 066102.

[Yan+18a] Jinfeng Yang et al. 'Structural correlation between communities and core-periphery structures in social networks: Evidence from Twitter data'. In: *Expert Systems with Applications* 111 (2018), pp. 91–99.

[Yan+18b] Jinfeng Yang et al. 'Structural correlation between communities and core-periphery structures in social networks: Evidence from Twitter data'. In: *Expert Systems with Applications* 111 (2018), pp. 91–99.

[YAT16] Zhao Yang, René Algesheimer and Claudio J Tessone. 'A comparative analysis of community detection algorithms on artificial networks'. In: *Scientific reports* 6.1 (2016), p. 30750.

[YZN11] Rong Yang, Leyla Zhuhadar and Olfa Nasraoui. ‘Bow-tie decomposition in directed graphs’. In: *14th International Conference on Information Fusion*. IEEE. 2011, pp. 1–5.

[Zac77] Wayne W Zachary. ‘An information flow model for conflict and fission in small groups’. In: *Journal of anthropological research* 33.4 (1977), pp. 452–473.

[ZMN15] Xiao Zhang, Travis Martin and Mark EJ Newman. ‘Identification of core-periphery structure in networks’. In: *Physical Review E* 91.3 (2015), p. 032803.

APPENDIX A

Appendix A of Thesis

A1 Data

For the data survey in Sec. 3.3, we consider all networks from the Netzschleuder repository Ref. [Pei] in Sec. 3.3(accessed Aug. 2022) which belong to one of 5 common domains (Online Social, Economic, Technological, Biological, Informational) and satisfy the following constraints: 1) number of nodes in the largest connected component is between 50 and $20,000$; 2) not bipartite; and 3) not multilayer and no multiple types of edges. For the list of all networks see Table. A.1. In order to allow the systematic application of the different community-detection methods, we removed all self-loops and we consider only the largest connected component of the network. We found 52 networks with 26 undirected graphs and 26 directed graphs.

The data for the case study in Sec. 3.4.1 is from Ref. [AG05] and corresponds to polblogs in the Netzschleuder repository. The data for the case study in Sec. 3.4.2 and Sec. 5.4.1 was obtained from the online platform X as described in Sec. 5.2. This data is available in our repository Ref. [Liu23]. The data for the case studies in Sec. 5.4.2 and Sec. 5.4.3 is from Ref. [Gai+21].

Type	**Name**	**N**	**E**	**Information/Edge Direction**
Undirected	facebook_friends	329	1954	–
	ego_social/gplus_10032969 8645326486178	2211	79585	–
	facebook_organizations/L1	5793	45266	–
	facebook_organizations/L2	1429	32876	–
	facebook_organizations/M1	3862	87324	–
	facebook_organizations/M2	320	2369	–
	facebook_organizations/S1	165	726	–
	facebook_organizations/S2	5524	94218	–
	blumenau_drug	75	181	–
	celegans_metabolic	453	4574	–
	fullerene_structures/C1500	1500	2250	–
	human_brains/BNU100258 64_1_DTI_AAL	116	1563	–
	interactome_pdz	161	209	–
	budapest_connectome/ all_20k	1015	70654	–
	celegans_interactomes/ wi2007	1108	1500	–
	collins_yeast	1004	8319	–
	interactome_vidal	2783	6007	–
	interactome_yeast	1458	1948	–
	kegg_metabolic/aae	880	2368	–
	tree-of-life/394	786	3874	–
	power	4941	6594	–
	route_views/19971108	3015	5156	–
	adjnoun	112	425	–

	polbooks	105	441	–
	wiki_science	677	6517	–
	bible_nouns	1707	9059	–
Directed	anybeat	12645	67053	Different
	qa_user/mathoverflow_c2a	13599	139534	Same
	inploid	14360	57101	Same
	faculty_hiring_us/domain _natural_sciences	2142	30381	Same
	faculty_hiring/business	113	8539	Same
	faculty_hiring/computer _science	206	4633	Same
	faculty_hiring/history	145	4347	Same
	un_migrations	232	11228	Same
	celegansneural	297	2359	Same
	cintestinalis	205	2854	Same
	foodweb_little_rock	183	2476	Same
	fresh_websAkatoreA	84	227	Same
	interactome_figeys	2217	6438	Different
	interactome_stelzl	1615	6075	Different
	messal_shale	700	6414	Different
	yeast_transcription	664	1066	Same
	ecoli_transcription/v1.0	329	456	Same
	celegans_2019/male_ chem-ical	559	5246	Same
	caida_as/20070917	8020	36406	Different
	gnutella/04	10876	39994	Different
	jung	6120	138706	Different
	software_dependencies/jdk	6434	150985	Different
	polblogs	1222	19086	Different

	webk/bwebkb_wisconsin_ link1	280	1106	Same
	word_adjacency/darwin	7377	46279	Same
	dblp_cite	12494	49687	Different

TABLE A.1. Datasets Table: 26 undirected graphs and 26 directed graphs. Columns include network type(Type), network name(Name),node number(N), edge number(E) and the consistency between information flow and edge direction(Information/Edge Direction).

A2 Robustness Estimation

We test the statistical robustness of our classification of communities in types by (1) comparing to a null model and (2) considering small perturbations of the underlying network.

A2.1 Null model

We consider a null model of random density matrix ω^{null} to assess the significance of the results we obtained applying our pairwise classification of community types to ω. Since our pairwise classification depends only on the ranking order of the entries in ω, we build ω^{null} as a random $B \times B$ matrix with entries from 1 to B^2. We then apply our classification – defined in Fig. 3.2 of the manuscript – to obtain the interaction classification matrix M^{null}, the structure type fraction P_{τ}^{null} from Eqs. (3.5), and the dominant type τ_d^{null} and occurring types τ_o^{null} – from Eqs. (3.6) and (3.7) of the manuscript– for the null model. The result obtained from this analysis represent our random expectation for an arbitrary partition of the network g in B communities. We further propose a statistic over many networks $\mathbf{g}$ by generating null models for each g with the same B obtained in the original analysis. This method is summarized in Algorithm 1. The dominance and occurrence obtained using this null model are reported in Fig. 3.4 of the manuscript.

Algorithm 1 Density Null Model

Require: number of communities B of network g
1: Randomly assign ranking from 1 to B^2 to $\omega^{\text{null}}_{B\times B}$
2: Compute M^{null} from the classification in Fig. 3.2
3: Calculate P^{null}_{τ} for community type τ from Eq. (3.5)
4: Get dominant type τ^{null}_{d} and occurring types τ^{null}_{o}
5: Repeat the steps above for many networks g to compute Dominance$(\tau)^{\text{null}}$ and Occurrence$(\tau)^{\text{null}}$ from Eq. (3.6) and (3.7)

A2.2 Robustness of categorical classification

Our community structure classification is computed for a given partition of the network. In practice, small variations of the network or algorithms (e.g., the initialization of optimization steps) may result in (slightly) different partitions. To have a better understanding of the robustness of our classification and results to such variations, we introduce a bootstrapping method for estimating the uncertainty of the results obtained from our classification.

Algorithm 2 Bootstrapping Robustness Estimation

Require: Graph g, partition p
1: Create a rewiring graph g'
2: Compute density matrix ω'
3: Get M' using classification in Eqs. (3.3) or (3.4)
4: Repeat above procedure k times
5: Calculate certainty P_{rs} from Eqs. (A.1)

Given a certain graph G (with N nodes and E edges) and partition p with its corresponding pairwise community classification matrix M, we first rewire the graph by randomly choosing E edges (with replacement) from the edge list of G. We denote this generated graph G'. Then we compute the density matrix ω' (based on G' and p) and we obtain a new interaction classification matrix M'. We repeat this procedure k times and compute the certainty for each pair of communities r and s, P_{rs}, as the fraction of all k interaction classifications that are equal to the original type

$$P_{rs} = \frac{1}{k}\sum_{1}^{k} \delta_{M_{rs},M'_{rs}}. \tag{A.1}$$

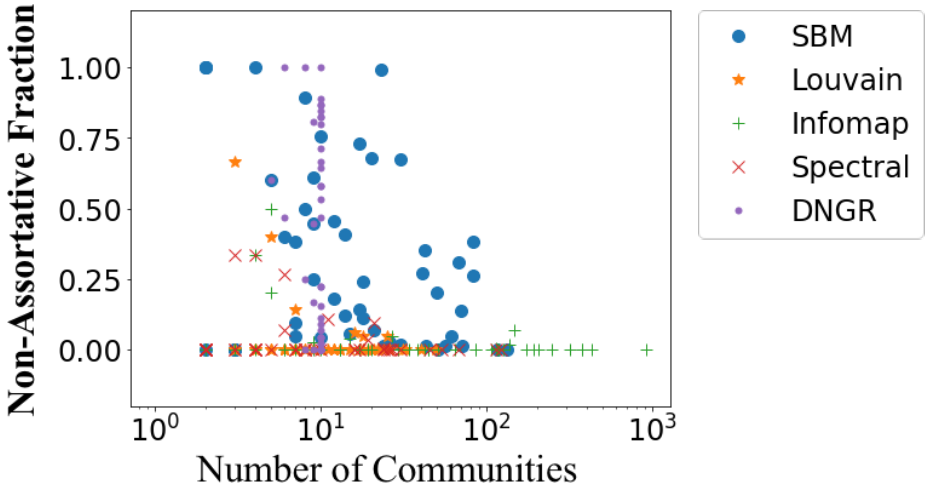


FIGURE A.1. Fraction of non-Assortative relationships (y-axis) as a function of the number of communities(x-axis) in empirical graphs. Each symbol represents a partition result of a network obtained using a specific community-detection method (see legend).

We applied this analysis in several cases and confirmed that our main conclusions remain unchanged. In particular, for the two case studies reported in Sec. 3.4, we find that all classifications for $B = 5$ lead to $P_{rs} > 0.99$ except for one core-periphery classification in the blogs network which had $P_{r,s} = 0.86$ (see Fig. A.2).

A3 Survey

Apart from dominance and occurrence shown in Sec. 3.2.3 of the main manuscript, we also consider an intermediate quantity – denoted proportional occurrence – to measure the classification result in empirical networks. Instead of focusing on the appearance of a type τ or in the most dominant type τ, here we focus on the fraction P_τ of pairs of a given type τ in each network. For each community structure type $\tau \in \{$"Assortative", "Core-Periphery", "Disassortative", "Source-Basin"$\}$, we average the fraction P_τ over a set of networks as

$$\text{Proportional Occurrence}(\tau) = \frac{1}{|\mathrm{g}|} \sum_{g} \sum_{\tau' \in \tau_o} P_{\tau'} \delta_{\tau,\tau'}. \tag{A.2}$$

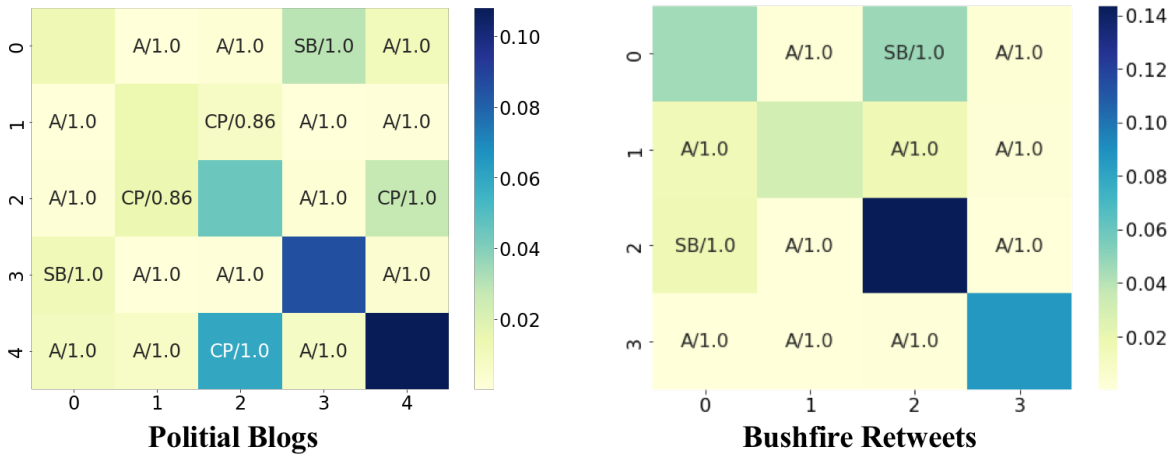


FIGURE A.2. Density matrices for the partition obtained by SBM with $B = 5$ for political blogs (left) and $B = 4$ for bushfire retweets(right). Colours indicate the community densities $\omega_{r,s}$. The blocks $b = 0, 1, \ldots, 4$ are ranked according to community size $N_r = \{354, 291, 291, 210, 76\}$(left) and $N_r = \{1473, 1031, 956, 563\}$(right) from community 0 to community 4(left)/3(right). As noted in the main text, the edge direction of the political blogs network is reversed when compared to the original dataset in order to fit the convention used in our text. The labels in each entry of the matrix correspond to the structure type τ and robustness score obtained from Eqs.(A.1) (see Sec. A2, Algorithm 2), where "A" stands for "Assortative", "CP" stands for "Core-Periphery", "D" stands for "Disassortative" and "SB" stands for "Source-Basin". Left: Community 0 and 3 are in a Source-Basin relationship. Communities 4 and 2 are in a Core-Periphery relationship and communities 2 and 1 are in a Core-Periphery relationship (i.e., community 2 is a periphery in respect to 4 and core in respect to 1). Right: Community 0,1,2 are climate change supporter groups. Community 0 and 2 are in a Source-Basin relationship.

This section shows additional analyses of our survey results. The results obtained in our survey using the proportional occurrence method are shown in Fig. A.3. Separated results for directed and undirected networks are shown in Fig. A.4. Finally, the results obtained using the DNGR method with different number of communities B are shown in Fig. A.5.

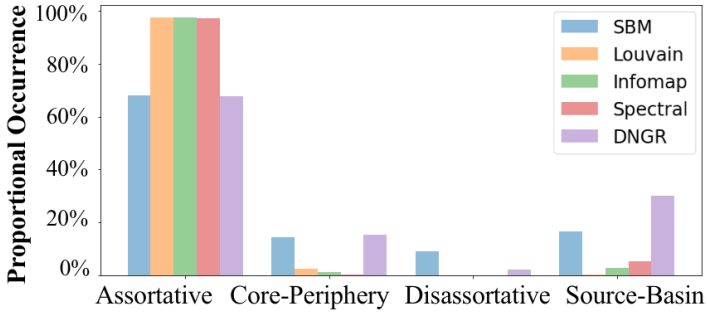


FIGURE A.3. Average fraction of pairs of the different types τ (Assortative, Core-Periphery, Disassortative, Source-Basin). For each network, we compute the proportional occurrence of type τ_i defined as Eq.(A.2). The reported results correspond to the average over 52 networks.

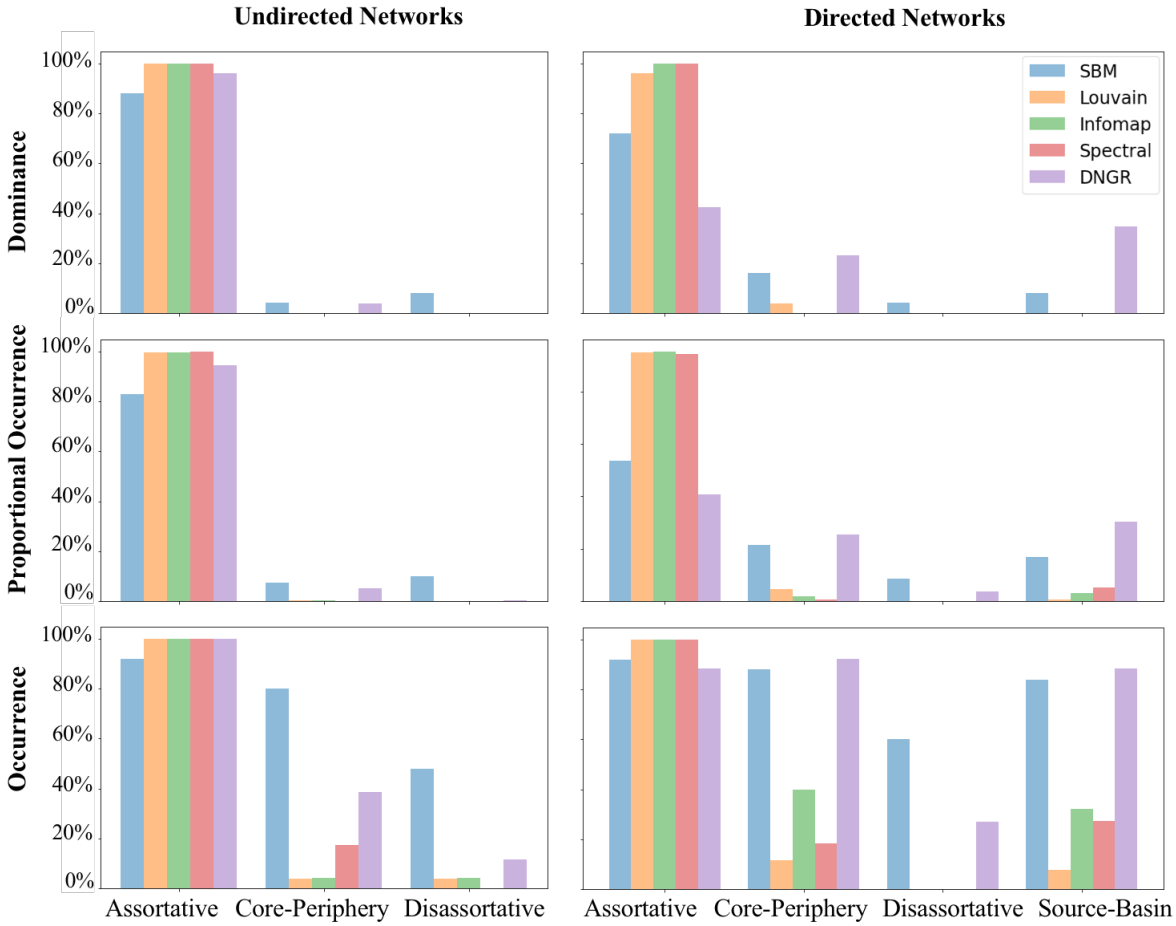


FIGURE A.4. Community structure type in empirical undirected graphs (left) and directed graphs (right).Top: fraction of networks (y-axis) in which community type (x-axis) is dominant. Middle: weighted fraction of networks in which community type appears. Bottom: fraction of networks in which community type appears.

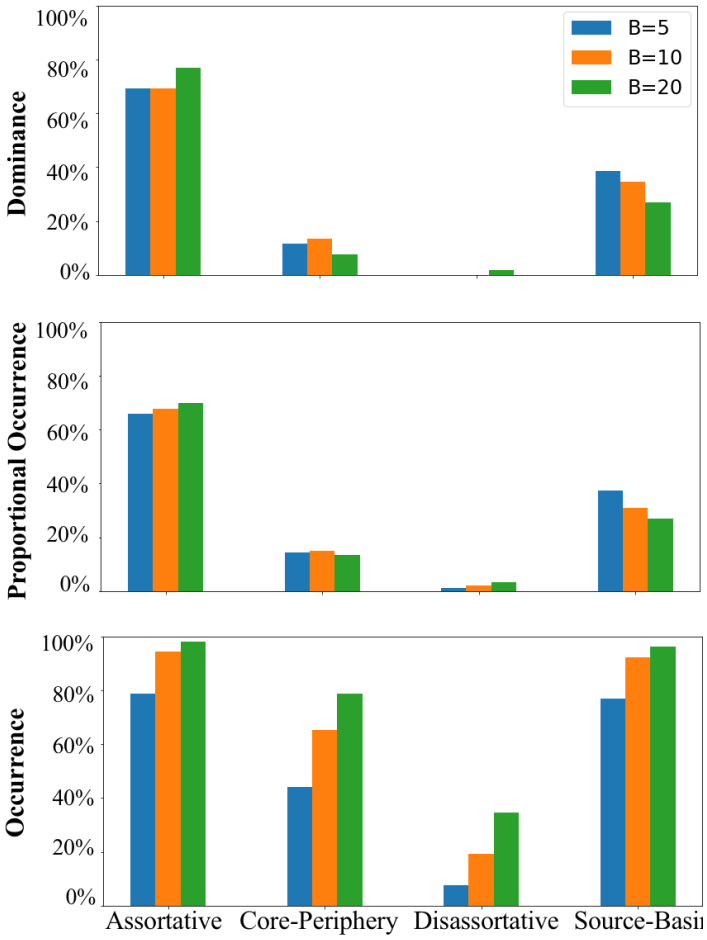


FIGURE A.5. Community structure type obtained applying the DNGR method to the empirical graphs in our survey. Results for three pre-defined number of communities are shown $B = 5, 10, 20$. Overall, the results with different B suggest an increasing occurrence of non-assortative structures and an increasing dominance (and proportional occurrence) of Assortative structure as B grows.

APPENDIX B

Appendix B of Thesis

B1 Derivation of General Solution

In this appendix we show the steps in the demonstration of Eq. (4.5).We start by computing the probabilities $P^{\pm}$ of increasing $(+)$ and decreasing $(-)$ edges for each of the moves, as used in Eq. (4.4). To do so, we notice that the within group edges e_{rr} is decided in four possible edge updates:

- In swap move, an extra edge adds to e^t_{rr} in the case that an edge $k \mapsto i$ from group s to group r ($g(k) = s, g(i) = r$) rewires to an edge $j \mapsto i$ coming from group r ($g(j) = r$). This happens when 1) we pick a focal node i in group r with probability $\frac{N_r}{N}$; 2) the node i plays assortative game with probability $P^S_r P^A_r$; 3) we pick an existing edge $k \mapsto i$ that is from different group with probability $1 - \beta_r$; and 4) we pick an non-edge $j \mapsto i$ that has $g(j) = r$ with probability $\frac{N_r}{N}$. These 4 steps are necessary and independent of each other. Therefore, the probability of gaining an edge within group r is calculated as the product of their probabilities as

$$P^+(\text{swap}) = \frac{N_r}{N} P^S_r P^A_r (1 - \beta_r) \frac{N_r}{N}. \tag{B.1}$$

- In swap move, $e_r r^t$ decreases one edge in the case that an edge $k \mapsto i$ from group r to group r ($g(k) = g(i) = r$) rewires to edge $j \mapsto i$ coming from group s ($g(j) = s$). It happens when 1) we pick a focal node i in group r with probability $\frac{N_r}{N}$; 2) the node i plays disassortative game with probability $P^S_r(1 - P^A_r)$; 3) we pick an existing edge $k \mapsto i$ that is from same group with probability β_r; and 4) we pick an non-edge

$j \mapsto i$ that has $g(j) \neq r$ with probability $\frac{N_s}{N}$. As above, the probability of losing an edge within group r is calculated as

$$P^-(\text{swap}) = \frac{N_r}{N} P_r^S (1 - P_r^A) \beta_r \frac{N_s}{N}. \tag{B.2}$$

- In change move, e_{rr}^t increases by 1 when the add move is chosen and a new edge $j \mapsto i$ coming from group r is added at random. It happens when 1) we pick a focal node i in group r with probability $\frac{N_r}{N}$; 2) node i plays add move with probability $(1 - P_r^S)\frac{1}{2}$; and 3) we pick a non-existing edge to add with approximate probability $\frac{N_r}{N}$. The probability in this case is thus

$$P^+(\text{change}) = (1 - P_r^S)\frac{1}{2}\frac{N_r}{N}\frac{N_r}{N}. \tag{B.3}$$

- In change move, e_{rr}^t decreases when an edge $k \mapsto i$ from group r is deleted. It happens when 1) we pick a focal node i in group r with probability $\frac{N_r}{N}$; 2) node i plays remove move with probability $(1 - P_r^S)\frac{1}{2}$; and 3) for each existing incoming edge from group r, we remove it with probability $\alpha_r \beta_r$. The probability in this case is thus

$$P^-(\text{change}) = (1 - P_r^S)\frac{1}{2}\frac{N_r}{N}\beta_r \alpha_r. \tag{B.4}$$

Introducing Eqs. (B.1)-(B.4) in Eq. (4.4), and using the definition of β_r in Eq. (4.3), we obtain

$$\begin{aligned} e_{rr}(t + \Delta t) = e_{rr}(t) + \frac{N_r}{N} P_r^S P_r^A \left(1 - \frac{e_{rr}(t)}{e_{rr}(t) + e_{sr}(t)}\right)\frac{N_r}{N} - \frac{N_r}{N} P_r^S (1 - P_r^A)\frac{e_{rr}(t)}{e_{rr}(t) + e_{sr}(t)}\frac{N_s}{N} \\ + (1 - P_r^S)\frac{1}{2}\frac{N_r}{N}\frac{N_r}{N} - (1 - P_r^S)\frac{1}{2}\frac{N_r}{N}\frac{e_{rr}(t)}{e_{rr}(t) + e_{sr}(t)}\alpha_r \frac{1}{\alpha_r} \end{aligned} \tag{B.5}$$

which can be simplified to Equation (4.5).

B2 Varying probability of remove move

In the evolution model introduced in Sec. 4.2.2, we chose the probability of choosing remove or add move to be $\frac{1}{2}$. In this section, we show that this choice is not restricting the cases obtained in our model because varying this probability our density matrix can be made

invariant with a proper choice of other parameters. We define the probability of choosing remove move as P^R and the probability of add move as $1 - P^R$. This probability affects the average in-degree calculation, with Eq. (4.1) written as

$$\frac{d\langle z\rangle}{dt} = -\alpha\langle z\rangle P^R + (1 - P^R), \tag{B.6}$$

with the corresponding fixed point solution

$$\langle z^*\rangle = \frac{1 - P^R}{\alpha P^R}. \tag{B.7}$$

With different P^R, we can choose new $\widetilde{\alpha} = \frac{\alpha P^R}{1-P^R}$ so that Eq. (4.2) still holds.

The proportion of assortative edges β_r is also affected by P^R, with Eq. (4.6) generalized to

$$\beta_r = \frac{P_r^S P_r^A + (1 - P_r^S)(1 - P_r^R)}{P_r^S(P_r^A + (1 - P_r^A)\frac{N_s}{N_r}) + (1 - P_r^S)(1 - P_r^R)\frac{N}{N_r}}, \tag{B.8}$$

By choosing $\widetilde{P_r^S} = \frac{P_r^S}{2(1-P_r^S)(1-P_r^R)+P_r^S}$, we can also get the same equation as Eq. (4.6).

B3 Computation of P^{S*}

Here, we calculate that in the equilibrium of our model, there is a finite value $P^{S*} > 0$ such that for any probability of swap move $P^S < P^{S*}$ only SB community types are obtained. We compute P^{S*} by considering the conditions for which no A,D, and CP can be obtained in our model, that is, the boundary between *SB* and other types in Fig. 4.5 is outside the probability space of $P_0^A, P_1^A \in [0, 1]$. We check the conditions of A,D and CP structures from Eqs. (4.6), (4.9) and (3.4) with assumption of $P_0^S = P_1^S$ and $b < 1$.

(1) Assortative structure can not be obtained if the transition from SB to A ($\omega_{01} > \omega_{00}$ to $\omega_{01} < \omega_{00}$) can not appear. The boundary (line 1 in Fig. 4.5) exceeds the probability space if $P_1^A > 1$ when $P_0^A = 1$. The largest probability of swap move that does not show A structure is then $\omega_{01} = \omega_{00}$ with $P_0^A, P_1^A = 1$ which we denote as $P^{S*}(A)$

and is computed as the solution (for x) of the implicit equation

$$\frac{(1-b)(1+x)(1+c-x+cx)}{2x(1+c-x+cx+b(1-c)(1+x))} = 1. \tag{B.9}$$

(2) Disassortative structure can not be obtained if the boundary between D and SB (line 2 in Fig. 4.5) exceeds the probability space, that is, boundary function $P_1^A < 0$ when $P_0^A = 0$. The largest probability of swap move that does not show D structure is then $\omega_{10} = \omega_{11}$ with $P_0^A, P_1^A = 0$ which we denote as $P^{S*}(D)$ and is computed as the solution (for x) of the implicit equation

$$\frac{bcx^2 + \frac{1}{2}x(1-x)(b+2bc-1) + \frac{1}{4}(1-x)^2(bc+b-c-1)}{x^2(1-b+bc) + \frac{1}{2}x(1-x)(c+1+bc+b)} = 0. \tag{B.10}$$

(3) Core-periphery structure can not be obtained if the boundary between CP and SB exceeds the probability space. Since we have two cases of CP with 1). group 1 is core and group 0 is periphery with boundary shown as line 3 in Fig. 4.5 and 2). group 0 is core and group 1 is periphery with boundary shown as line 4 in Fig. 4.5. The conditions of no CP would be $P_1^A > 1$ when $P_0^A = 0$ for the first case and $P_1^A < 0$ when $P_0^A = 1$ for the second case. The largest probability of swap move that does not show first case of CP is $\omega_{01} = \omega_{10}$ with $P_0^A = 0, P_1^A = 1$ which we denote as $P^{S*}(\text{first CP})$ and is computed as the solution (for x) of the implicit equation

$$\frac{(1-bc)x^2 + \frac{1}{2}x(1-x)(1+\frac{1}{c+1}-\frac{b}{c+1}-bc) + \frac{1}{4}(1-x)^2(c+1-bc-b)}{x^2(1-bc+b) + \frac{1}{2}x(1-x)(c+1-bc^2+b-bc)} = 1. \tag{B.11}$$

The largest probability of swap move that does not show second case of CP is $\omega_{00} = \omega_{11}$ with $P_0^A = 1, P_1^A = 0$ which we denote as $P^{S*}(\text{second CP})$ and is computed as the solution (for x) of the implicit equation

$$\frac{bcx^2 + \frac{1}{2}x(1-x)(2bc+b-c) + \frac{1}{4}(1-x)^2(bc-c+b-1)}{x^2(c+bc-b) + \frac{1}{2}x(1-x)(c+1+bc-b)} = 0. \tag{B.12}$$

The value of P^{S*} is thus the minimum value among the above four values

$$P^{S*} = \min\{P^{S*}(A), P^{S*}(D), P^{S*}(\text{first CP}), P^{S*}(\text{second CP})\}. \tag{B.13}$$

B4 No SB or CP in modified density matrix

Instead of comparing with the number of all possible links, another way of measuring the density strength is dividing the number of links from group r to s by the product of all existing out links of group r and existing in links of group s, defined as

$$\omega_{rs} = \frac{e_{rs}}{e_r^{\text{out}} e_s^{\text{in}}} \tag{B.14}$$

where $e_r^{\text{out}} = e_{rr} + e_{rs}$ is the total out degrees of group r and $e_s^{\text{in}} = e_{ss} + e_{rs}$ is the total in degrees of group s.

This modified formula of density matrix takes degrees of nodes into account and thus removes the effect of degree difference between groups. We notice that neither SB or CP structures that requires different degrees between groups can be found with new density matrix.

We first show the conditions of SB structure are mutually conflicting. The case of group r as source and group s as basin requires the following conditions

$$\omega_{rr} > \omega_{rs} \Rightarrow \frac{e_{rr}}{(e_{rr}+e_{sr})(e_{rr}+e_{rs})} > \frac{e_{rs}}{(e_{ss}+e_{rs})(e_{rr}+e_{rs})} \Rightarrow e_{rr}e_{ss} > e_{rs}e_{sr} \tag{B.15}$$

$$\omega_{sr} > \omega_{ss} \Rightarrow \frac{e_{sr}}{(e_{ss}+e_{sr})(e_{rr}+e_{sr})} > \frac{e_{ss}}{(e_{ss}+e_{sr})(e_{ss}+e_{rs})} \Rightarrow e_{rr}e_{ss} < e_{rs}e_{sr} \tag{B.16}$$

The conditions of CP structure are also mutually conflicting with following inequalities for case of group r as core and group s as periphery

$$\omega_{rr} > \omega_{sr} \Rightarrow \frac{e_{rr}}{(e_{rr}+e_{sr})(e_{rr}+e_{rs})} > \frac{e_{sr}}{(e_{ss}+e_{sr})(e_{rr}+e_{sr})} \Rightarrow e_{rr}e_{ss} > e_{rs}e_{sr} \tag{B.17}$$

$$\omega_{rs} > \omega_{ss} \Rightarrow \frac{e_{rs}}{(e_{rr}+e_{rs})(e_{ss}+e_{rs})} > \frac{e_{ss}}{(e_{ss}+e_{sr})(e_{ss}+e_{rs})} \Rightarrow e_{rr}e_{ss} < e_{rs}e_{sr} \tag{B.18}$$

The above conditions show that CP and SB could not be observed in two groups setting using modified density matrix which takes degrees into account.

Algorithm 3 Description of the algorithm used to implement our model of network evolution.

Initialization:: Erdős-Rényi directed network with N nodes and $2/N < q < 1 - 2/N$ edge creation probability

1.: Randomly choose one node $i \in [1, N]$.

2.: Choose $u \in [0, 1]$ uniformly at random and decide update:

(Swap) If $u < P^G(g(i))$:
Randomly pick k such that $A_{ki} = 1$ and j such that $A_{ji} = 0$
If $g(k) = g(j)$: update ($A_{ki} \leftarrow 1$ and $A_{ji} \leftarrow 0$) or ($A_{ki} \leftarrow 0$ and $A_{ji} \leftarrow 1$) by chance.
If $g(k) \neq g(j)$: choose $v \in [0, 1]$ and:
(A) If $v < P^A(g(i))$ (assortative):
if $g(k) = g(i)$: update ($A_{ki} \leftarrow 1$ and $A_{ji} \leftarrow 0$)
if $g(j) = g(i)$: update ($A_{ki} \leftarrow 0$ and $A_{ji} \leftarrow 1$)
(D) If $v \geq P^A(g(i))$ (disassortative):
if $g(k) = g(i)$: update ($A_{ki} \leftarrow 0$ and $A_{ji} \leftarrow 1$)
if $g(j) = g(i)$: update ($A_{ki} \leftarrow 1$ and $A_{ji} \leftarrow 0$)

(Change) If $u \geq P^G(g(i))$:
Choose $v \in [0, 1]$ uniformly at random:
(R) If $v < 0.5$ (remove):
update $A_{ki} \leftarrow 0$ with probability $\alpha(g(i))$ for all k such that $A_{ki} = 1$.
(A) If $v \geq 0.5$ (add):
update $A_{ki} \leftarrow 1$ for a randomly picked k such that $A_{ki} = 0$.

3.: Update $t \mapsto t + \Delta t$; go to Step 1.

APPENDIX C

Appendix C of Thesis

C1 Community Detection

We use degree-corrected hierarchical Stochastic Block Model(hSBM) as described in Sec. 2.3.4 to find communities of users in all networks of three datasets in Chapter 5. However, the hSBM (as well as other community detection methods) has a shortcut that the exact solution is computationally intractable and the application on empirical networks relies on approximations of optimum such that multiple partitions can be found with similar score on objective function [Pei21]. The susceptible answers of community detection methods make the interpretation of networks based on communities unreliable. In order to increase the robustness of SBM result, we proposed a step-wise methodology including taking consensus partitions, applying Metropolis-Hastings acceptance-rejection Markov chain Monte Carlo(MCMC) to find optimal partition and extracting "stable" states as final result.

Consensus Partition One way to reduce the uncertainty is collectively analyzing many outputs of community detection methods. We use the method proposed by Peixoto of taking many different partitions and recovering one single partition that most results are aligned with. We call this recovered partition a consensus partition. The advantage of this method is that it provides us with a more robust answer than a single particular one.

Markov chain Monte Carlo(MCMC) While consensus partition provides a compact description of heterogeneous partition results, it does not guarantee an optimal solution that has a minimum description length in the hSBM model. To further improve our result, we perform an extra step with a Metropolis-Hastings acceptance-rejection MCMC to find the solution

that has smaller and smaller description length and in the long enough time [Pei14a], find the optimal solution. The process we apply MCMC is as follows

(1) Start from consensus partition
(2) Apply MCMC for 10000 times
(3) Record the 10000th partition and its description length
(4) Repeat steps (2),(3)
(5) Stop when the improvement on description length is small

C2 Community Information

For each network of three datasets in Sec. 5.2, we apply the hSBM model for 10 iterations and find consensus partition from these 10 realizations. For multilayer bushfire network, we take an extra MCMC step for 75 iterations to find optimal partition (see Fig. C.1). We focus on top 3 levels with at least two groups detected. The community size is listed in Table. C.1. Partition overlapping between single layer network (retweet-only and reply-only) and multilayer network (retweet and reply) is shown in Table. C.3. Assortativity of all networks is measured by three methods: Modularity, Assortative fraction, and Diagonal Proportion, and the detail information is listed in Table. C.2. The density matrices for partitions of bushfire retweet network and NYE retweet, reply, multilayer networks are shown in Fig. C.2. Fig. C.3, and Fig. C.4. The abstract merging process of hierarchical partitions is shown in Fig. C.5 for the election dataset and Fig. C.6 for the NYE dataset.

C3 User Feature

Twitter API provides more data than just tweet text and retweet/reply activity records. For the bushfire data used in Sec. 3.4.2 and Sec. 5.4.1, the profile of all users was provided by the co-supervisor A/Prof. Tristram Alexander. The user metadata includes many features, such as verified account and location, providing further insights into the group's role in this bushfire

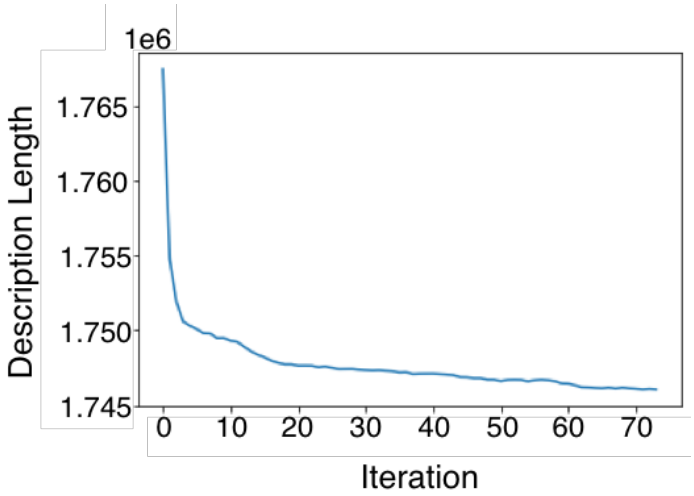


FIGURE C.1. Description length decrease during MCMC iterations for multilayer bushfire network. We start from the consensus partition obtained from 10 realizations of hSBM partitions and perform MCMC for 75 iterations. The description length of partitions decreased for the first 62 iterations and then became stable.

Network	Bushfire		
Interaction	retweet	reply	multilayer
Level 1	2	2	2
Level 2	4	8	4
Level 3	9	21	8
Network	Election		
Interaction	retweet	reply	multilayer
Level 1	2	2	2
Level 2	3	3	5
Level 3	7	9	16
Network	NYE		
Interaction	retweet	reply	multilayer
Level 1	2	2	3
Level 2	4	6	9
Level 3	10	16	27

TABLE C.1. Community size information of hSBM model for bushfire, election and NYE datasets detected by hSBM. The consensus partition is found from 10 realizations of hSBM results for each network (except multilayer bushfire network where the 70th partition result from MCMC is used).

discussion. The description of 5 user features used in this project is listed in Table. C.4. The feature information of groups of users identified by the hSBM model is listed in Table. C.5.

Network	Bushfire			
Interaction	retweet	reply	multilayer-retweet	multilayer-reply
Community Size	3	4	4	4
Modularity	0.18	0.09	0.31	0.27
Assortativity	1	0.17	0.83	0.67
Diagonal Proportion	0.85	0.33	0.69	0.51
Network	Election			
Interaction	retweet	reply	multilayer-retweet	multilayer-reply
Community Size	7	3	5	5
Modularity	0.37	0.24	0.44	0.33
Assortativity	0.57	1.0	0.8	0.9
Diagonal Proportion	0.10	0.90	0.75	0.61
Network	NYE			
Interaction	retweet	reply	multilayer-retweet	multilayer-reply
Community Size	4	6	3	3
Modularity	0.31	0.44	0.24	0.33
Assortativity	0.67	0.93	0.67	0.72
Diagonal Proportion	0.70	0.72	0.35	0.57

TABLE C.2. Assortativity measured in modularity, assortativity fraction and diagonal proportion for three datasets. Assortativity is computed by Eqn. 3.5. The diagonal proportion is computed by the inter-density divided by intra-density. The community information used for each network is from one level of hSBM model and the exact size is included in the table.

	Climate Multilayer	Election Multilayer	NYE Multilayer
Retweet	0.43	0.38	0.46
Reply	0.004	0.01	0.12

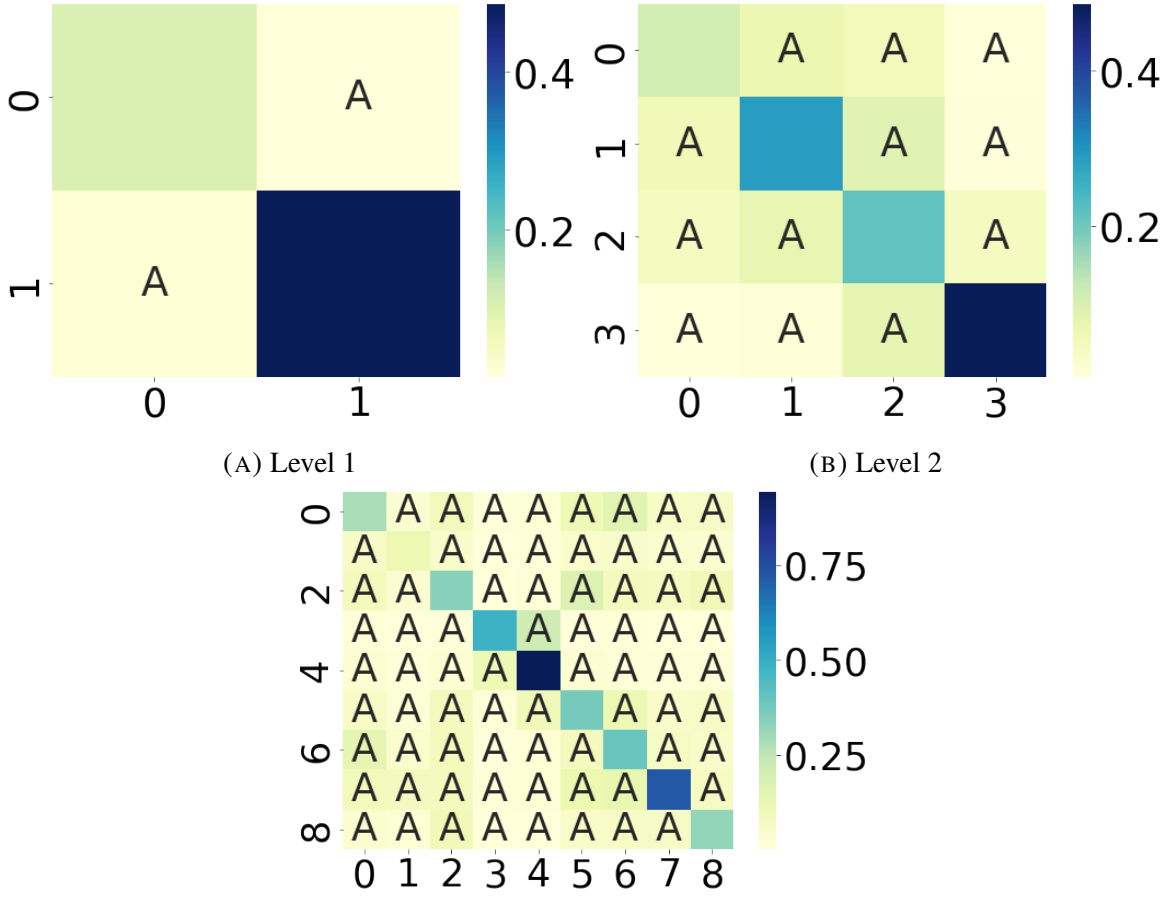
TABLE C.3. Overlapping between partitions from single layer networks and partitions from multilayer networks for three datasets. The overlapping is calculated by Normalised Mutual Information (NMI). The table shows partitions found by multilayer networks have around 40% overlapping with one found by retweet networks, and have small overlapping with one found by reply networks. The reason why retweet has a higher similarity is that three datasets have more retweets than replies, and thus the community detection is influenced more by retweet interactions.

Feature	Description
verified account	if the user's account is verified or not
followers count	the count of user's followers
friends count	the count of user's friends
status count	the count of user's status(including tweet/retweet/reply)
Australian check	if the user is Australian account or not

TABLE C.4. Table of description of user features of bushfire data used in Sec. 3.4.2 and Sec. 5.4.1. The verified account, followers, friends and status count features are extracted from user profile in Twitter. The Australian check is based on the location information in user profile.

Retweet Bushfire Network					
Group	Verified Account	Followers	Friends	Status	Australian
0	30%	392954	4000	25158	45%
1	25%	32196	2909	50791	87%
2	27%	468215	3497	35442	80%
4	21%	383859	10167	47308	24%
Multilayer Bushfire Network					
Group	Verified Account	Followers	Friends	Status	Australian
0	22%	19493	2749	40158	90%
1	35%	636953	4571	29763	34%
2	15%	137950	7993	44132	44%
3	51%	875915	2402	42087	93%

TABLE C.5. Detailed information of user features of from user profile for bushfire datasets. The group information is from the second level of hSBM partitions on retweet network and multilayer network.

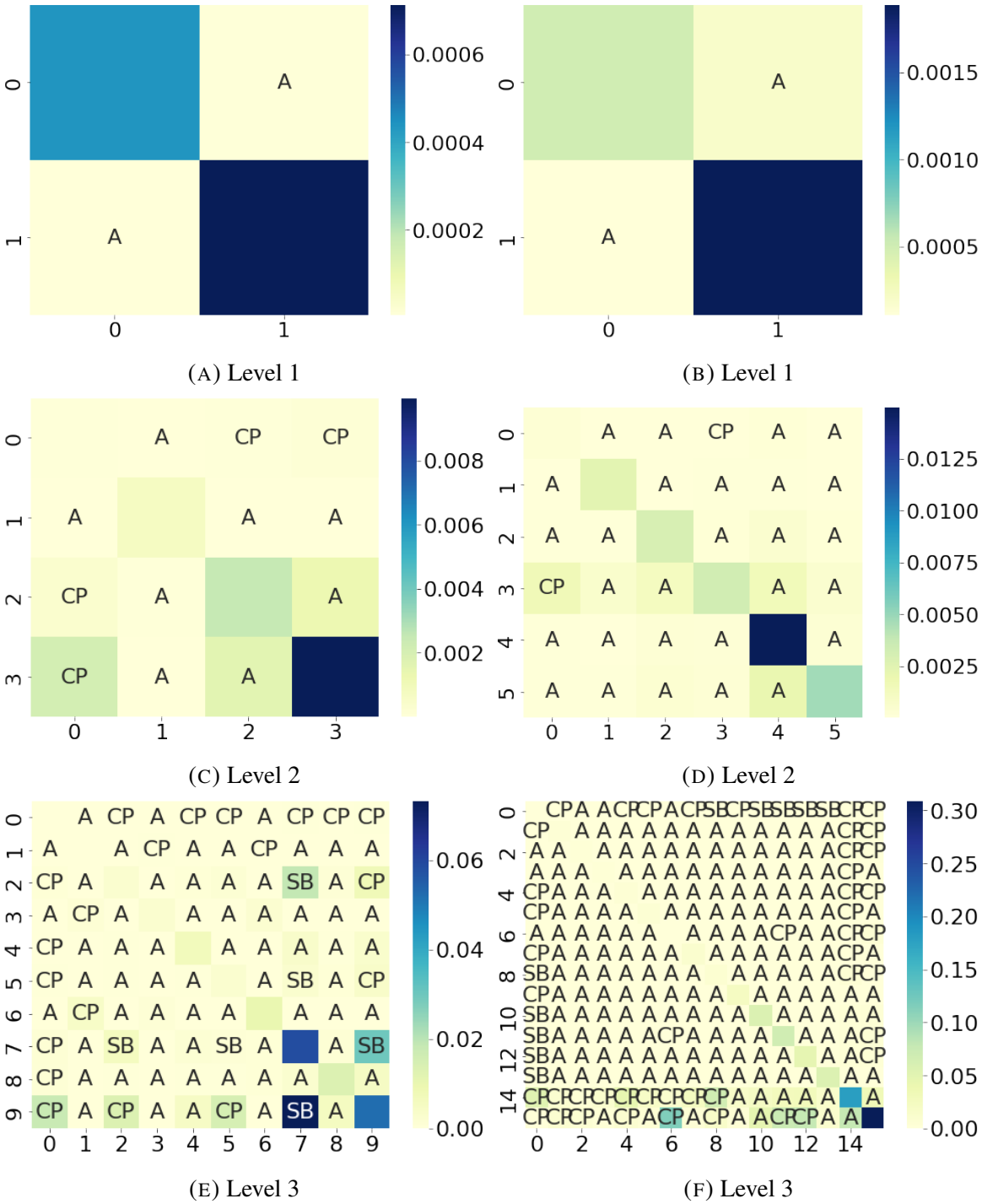


(A) Level 1

(B) Level 2

(C) Level 3

FIGURE C.2. Density matrices for the partition obtained by hSBM based on bushfire retweet network on level 1 (A), level 2 (B) and level 3 (C). The blocks are ranked according to community size $N_r = \{3671, 358\}$ (A), $N_r = \{1740, 1248, 683, 358\}$ (B), $N_r = \{1248, 1050, 621, 358, 225, 182, 178, 98, 69\}$ (C). The labels in each entry of the matrix correspond to the structure type τ where "A" stands for "Assortative", "CP" stands for "Core-Periphery", "D" stands for "Disassortative" and "SB" stands for "Source-Basin".



(A) Level 1

(B) Level 1

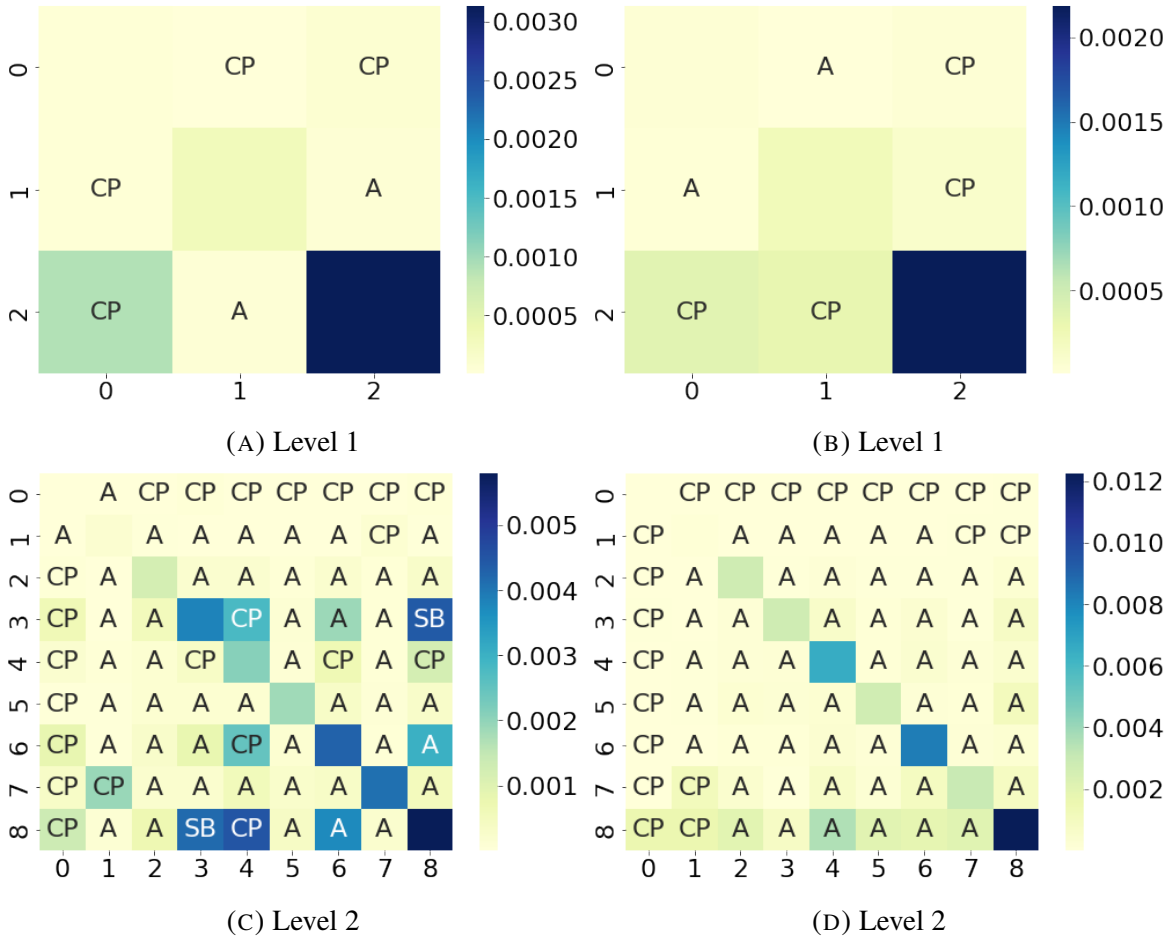
(C) Level 2

(D) Level 2

(E) Level 3

(F) Level 3

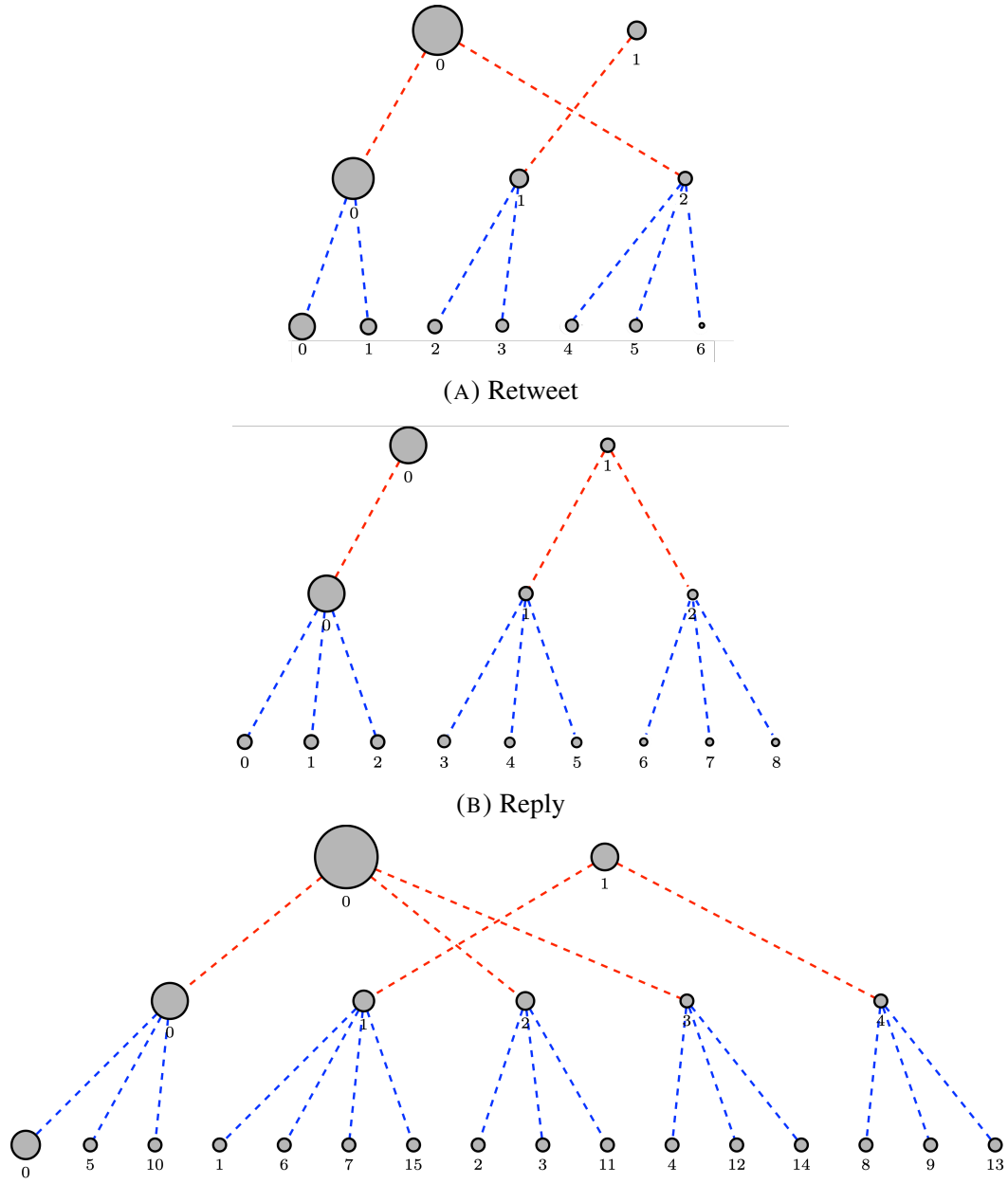
FIGURE C.3. Density matrices for the top 3 levels of partition obtained by hSBM based on NYE retweet network(Left) and reply network(Right). The blocks are ranked according to community size $N_r = \{6471, 2761\}$ (A), $N_r = \{5047, 2761, 744, 680\}$ (C), $N_r = \{5047, 2079, 496, 480, 376, 249, 202, 131, 119, 53\}$ (E), $N_r = \{4533, 1614\}$ (B), $N_r = \{3043, 919, 877, 613, 362, 333\}$ (D), $N_r = \{2226, 651, 593, 530, 464, 353, 322, 251, 226, 138, 95, 93, 91, 83, 20, 11\}$ (F). The labels in each entry of the matrix correspond to the structure type τ where "A" stands for "Assortative", "CP" stands for "Core-Periphery", "D" stands for "Disassortative" and "SB" stands for "Source-Basin".



(A) Level 1 (B) Level 1

(C) Level 2 (D) Level 2

FIGURE C.4. Density matrices for the top 2 levels of partition obtained by hSBM based on NYE multilayer network according to retweet interaction(Left) and reply interaction(Right). The blocks are ranked according to community size $N_r = \{8075, 4142, 1186\}$ (A,B), $N_r = \{6929, 3331, 687, 631, 459, 422, 367, 351, 188\}$ (C,D). The labels in each entry of the matrix correspond to the structure type τ where "A" stands for "Assortative", "CP" stands for "Core-Periphery", "D" stands for "Disassortative" and "SB" stands for "Source-Basin".

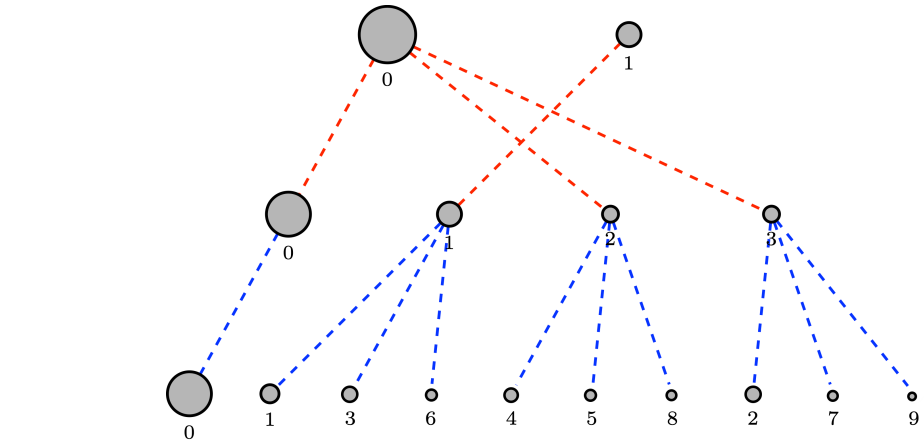


(A) Retweet

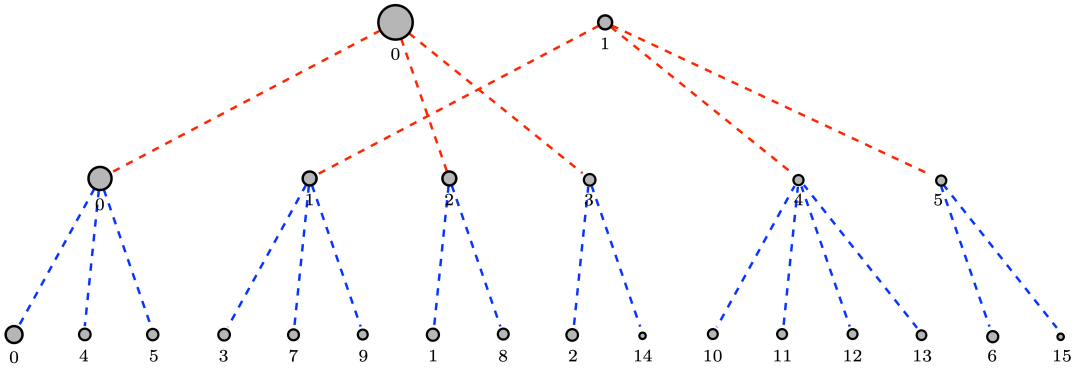
(B) Reply

(C) Multilayer

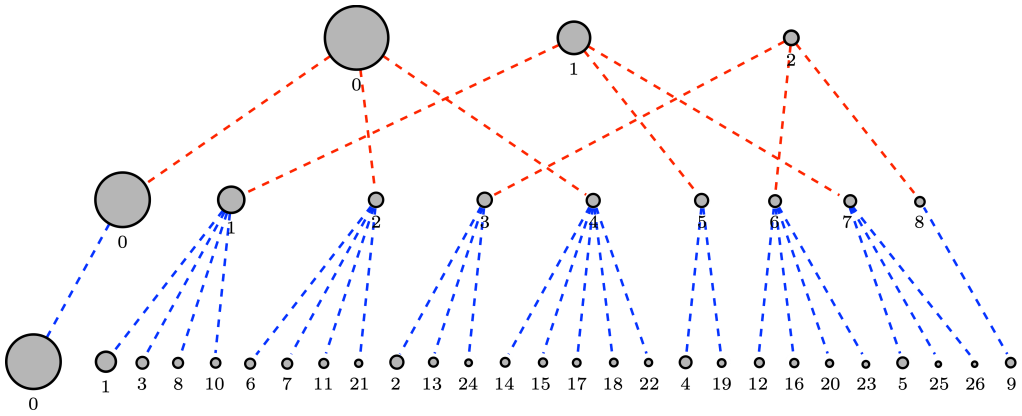
FIGURE C.5. Partitions from hSBM of election retweet network(TOP), reply network(Middle) and multilayer network(Bottom). We ignore the retweeting/replying interactions in each level and focus on the divide-and-conquer structure. Each node represents a community with size corresponding to community size listed in Fig. 5.9, 5.10.



(A) Retweet



(B) Reply



(C) Multilayer

FIGURE C.6. Partitions from hSBM model of NYE retweet network(TOP), reply network(Middle) and multilayer network(Bottom). We ignore the retweeting/replying interactions in each level and focus on the divide-and-conquer structure. Each node represents a community with size corresponding to community size listed in Fig. C.3, C.4.